A REVIEW ON PREDICTION OF MISSING DATA IN MULTIVARIABLE TIME SERIES

P. Vikas Kumar
PG Scholar
Department of CSE
Chaitanya Bharathi Institute of Technology, Hyderabad.
Email: perumalavikaskumar@gmail.com

M. Venu Gopalachari
Associate Professor
Department of CSE
Chaitanya Bharathi Institute of Technology, Hyderabad.
Email: venugopal.m@cbit.ac.in

ABSTRACT
Time series models help to track the behaviour for process or metrics over a set of time periods. The advancements in the complexity of the time series data leads to the evolution of multivariable time series analysis demanding qualitative machine learning techniques. Data missing in collections of multivariable time series occurs frequently is a challenging issue, due to the involvement of sensors to generate huge amount of data, which affects the quality of the data analysis. In order to overcome this issue, there is a need of applying efficient missing data prediction models. This paper addresses the impact of missing data in multivariable time series and presents the survey of the various models exists to handle missing data. The focus of this paper is on matrix factorization models in smoothing time series data and also the future directions to improve the quality of the missing data prediction.

KEYWORDS: - Time Series data, Missing Data Prediction, Matrix Factorization, Smoothing

INTRODUCTION

Time series data plays a major role in many applications such as Economic Forecasting, Sales Forecasting, Budgetary Analysis, Stock Market Analysis, Yield Projections, Process and Quality Control, Inventory Studies, Workload Projections, Utility Studies, Census Analysis etc., which involve data collected over a large time frame, and also a very large number of possibly interdependent time series [1]. A time series is a sequence of values that are linked to a timestamp, in other contexts sometimes it refers to a series with multiple values at the same timestamp is known as multivariable time series data [2]. For example, weather forecasting applications involve data collected from the huge number of sensors, over finite time periods continuously. Similarly, inventory tracking applications would collect data of the large set of items everyday for multiple years. In the “Internet of Things”, a large number of sensors are used to produce multivariable time series of the external environment, e.g., the air or water quality. The primary goal of time series forecasting is to predict the future trend of the data based on the historical records.

In general, time series data incurs missing values due to machine malfunctions, and/or simple human errors. The data available to build time series models are often characterized by missing values, due to various causes such as sensor faults, problems of not reacting experiments, not recovering work situations, transferring data to digital systems. The methodologies to predict missing data in time series strikes with two major challenges, one is Scalability in the presence of high-dimensional time series and the other one is the flexibility to handle missing values. One simple method to predict the missing data in time series model is linear interpolation [4]. There are several models exists in the literature to predict the missing values in multivariable time series [6]. These models include traditional approaches such as Tree based, vector based, regression based and graph based methodologies in machine learning.
Time series modelling mainly classified as parametric and non-parametric where parametric approach aim to fit a statistical model for time series data [19 slide]. On the other hand non parametric approaches will not have structural assumptions about the underlying structure of the process. [26 Slide] As the time series data itself is different from traditional structured data, the way of fitting a model needs complex pre-processing techniques such as smoothing, de-trending, auto correction etc. [7]. But the models with matrix factorization approach are relatively performing better in missing data prediction [10]. This survey focuses merits and demerits of the missing data prediction in multivariable time series models.

The Matrix Factorization (MF) plays an important role in the Collaborative Filtering recommender system. MF have recently received greater exposure, mainly as an unsupervised learning method for latent variable decomposition and dimensionality reduction [11] [12]. Prediction of ratings and Recommendations can be obtained by a wide range of algorithms, while Neighborhood-based Collaborative Filtering methods are simple and intuitive. The Matrix Factorization techniques are usually more effective because they allow use to discover the latent features underlying the interactions between users and items. Matrix Factorization is simply a mathematical tool for playing around with matrices, and is therefore applicable in many domains where one would like to find out something hidden under the data. Single Value Decomposition (SVD) and Principle Component Analysis (PCA) are well known Matrix Factorization models for identifying latent factors in the field of Information Retrieval to deal with Collaborative Filtering challenges [11][12][13].

**TIME SERIES DATA**

Time series data are a collection of ordered observations recorded at a specific time, for instance, hours, months, or years. Most often, the observations are made at regular time intervals. Time series analysis accounts for the fact that data points taken over time may have an internal structure, such as autocorrelation, trend or seasonal variation. There are two main goals of time series analysis. First, we hav 2 identify the nature of the phenomenon represented by the sequence of observations in the time series data. Second, we use the data to forecast or predict future values of the time series variable. Both of these goals require that we identify the pattern of observed time series data and more or less formally describe it. Usually, time series consists of four components: trend, seasonality, cyclic and random. The trend refers to the long term moment in a time series it indicates whether the observation values are increasing or decreasing over time. For example a steady increase in sales over month. Seasonal refers to the fixed fluctuations in time series observations during specific time periods. Cyclic is also the observation of time series fluctuation but it is not fixed over a time period. Random component is the series that does not fall under any of the previous category.

The pre-processing of the time series includes smoothing, detrending, filtering seasonality so that the series attains stationarity. A trend usually refers to a change in the mean over time. The process of removing trend form from time series data is known as detrending. When we detrend the time series data, it will remove an aspect from the data that we think is causing some kind of distortion.

A time series is stationary if it fluctuates around a constant mean. The non-stationary series includes a longer-term secular trend. However, this does not present an insolvable problem, since there are several methods which, in fact, allow us to transform a non-stationary series into one which is indeed stationary. A non-stationary time series usually includes a trend element. The stationarity of a time series can be authorized by three conditions, first one is mean of the variable for each time period must be constant, second one is the variance of the variable for each time period have to be approximately equal and the third one is the covariance of the adjacent time periods has to be same. Stationarity is basic assumption in the time series analysis. A time series is stationary if it fluctuates around a constant mean. The non-stationary series includes a longer-term secular trend. The advantage of dealing with stationary series is that their statistical properties are independent of time. It is important because without stationarity forecasting might become difficult. There are vari-
ants of stationarity based on the focus to achieve stationarity in the series namely strictly stationarity, weakly stationarity and co-variance stationary.

**Time series forecasting/Analysis**

There are many methods of time series model. The user's application and preference will decide the selection of the appropriate technique. Time series analysis comprises methods for analyzing time series data to extract meaningful statistics and other characteristics of time series data. It focuses on comparing values of a single time series and multiple dependent time series at different points in time. Time series forecasting is to predict future values of a time series based on previously observed values in the time series. The time series model can be classified into Filtering, Smoothing, and Decomposition categories.

Initially SVMs were considered to solve pattern classification problems such as optimal character recognition, face identification and text classification, etc. But soon they found wide applications in other domains such as function approximation, regression estimation and time series prediction problems. SVM techniques are based on the Structural Risk Minimization rule. The objective of SVM is to find a decision rule with good generalization capability through selecting some particular subset of training data called support vectors. In this method, an best possible separating hyper plane is constructed and after nonlinearly mapping the input space into a higher dimensional feature space. Thus, the quality and complexity of SVM solution does not depends directly on the input space. Another important characteristic of SVM is that the training process is equivalent to solving a linearly inhibited quadratic programming problem. In contrary to other networks training, the SVM solution is always exclusive and globally optimal. However a major difficulty of SVM is that when the training size is large, it requires an enormous amount of computation which increases the time complexity of the solution.

The time series models follow Box-Jenkins architecture which includes Auto Regressive, Moving Average, Auto Regressive and Moving Average, and Auto Regressive Integrated Moving Average models as shown in figure 1. The Box-Jenkins model assumes that the time series is stationary. Box and Jenkins recommend differencing non-stationary series one or more times to achieve stationarity. Doing so produces an ARIMA model, with the "I" standing for "Integrated". Some formulations transform the series by subtracting the mean of the series from each data point. This yields a series with a mean of zero. Whether you need to do this or not is dependent on the software you use to estimate the model. Box-Jenkins models can be extended to include seasonal autoregressive and seasonal moving average terms. Although this complicates the notation and mathematics of the model, the underlying concepts for seasonal autoregressive and seasonal moving average terms are similar to the non-seasonal autoregressive and moving average terms.

The most general Box-Jenkins model includes difference operators, autoregressive terms, moving average terms, seasonal difference operators, seasonal autoregressive terms, and seasonal moving average terms. As with modelling in general, however, only necessary terms should be included in the model. Rather than use present values of the forecast variable in a regression, a moving average model use past forecast errors in a regression-like model. But moving average models should not be mystified with moving average smoothing. A rousing average model is used for forecasting vision values while moving average smoothing is used for estimate the trend-cycle of past values.

Before forecasting we need to check the stationarity and apply the suitable model. There are three primary stages in building a box-Jenkins time series model: Identification, Estimation and Diagnostic checking.

**Filtering method**

The technique can be used to isolate frequencies that are of physical interest from those that are not. It can be used to remove high frequency noise or low frequency trends from time series and leave unaltered the frequencies of interest. These applications are called low-pass and high-pass filtering, respectively. A band-pass filter will remove both high frequencies and low frequencies...
and leave only frequencies in a band in the middle. Band-pass filters tend to make even noise look periodic, or at least quasi-periodic. We will begin by noting a few important theorems that constitute the fundamental tools of non-recursive filtering.

**Kalman filtering**

A Kalman filter is an optimal estimator which infers parameters of interest from indirect, inaccurate and uncertain observations. It is recursive so that new measurements can be processed as they arrive. Optimal in the sense if all noise is Gaussian, the Kalman filter minimises the mean square error of the estimated parameters. If the noise is not Gaussian, given only the mean and standard deviation of noise, the Kalman filter is the best linear estimator. Non-linear estimators may be better. It is so popular because: Due to optimality and structure it gives the good results. For online real time processing it is very convenient form. Easy to formulate and implement. Measurement equations not to be inverted.

**Smoothing method**

Smoothing is usually done to help us better see patterns, trends in time series. For seasonal data, we might smooth out the seasonality so that we can identify the trend. Smoothing doesn’t provide us with a model, but it can be a good first step in describing various components of the series. The term filter is sometimes used to describe a smoothing procedure. For instance, if the smoothed value for a particular time is calculated as a linear combination of observations for surrounding times, it might be said that we’ve applied a linear filter to the data (not the same as saying the result is a straight line, by the way). Exponential smoothing methods give larger weights to more recent observations, and the weights decrease exponentially as the observations become more distant. These methods are most effective when the parameters describing the time series are changing slowly over time. The Simple Exponential Smoothing method is used for forecasting a time series when there is no trend or seasonal pattern, but the mean (or level) of the time series y_t is slowly changing over time.

Decomposition procedures are used in time series to describe the trend and seasonal factors in a time series. It might also include long-run cycles, holiday effects, and day of week effects and so on. Here, we’ll only consider trend and seasonal decompositions. One of the main objectives for

![Figure 1: Interactive process of Box Jenkins Methodology.](image-url)
a decomposition is to estimate seasonal effects that can be used to create and present seasonally adjusted values. A seasonally adjusted value removes the seasonal effect from a value so that trends can be seen more clearly. For instance, in many regions of the U.S. unemployment tends to decrease in the summer due to increased employment in agricultural areas. Thus a drop in the unemployment rate in June compared to May doesn’t necessarily indicate that there’s a trend toward lower unemployment in the country. To see whether there is a real trend, we should adjust for the fact that unemployment is always lower in June than in May.

MISSING DATA IMPACT
Missing data or missing values is defined as the data value that is not stored for a variable in the observation of interest. The problem of missing data is relatively common in any applications. Missing data can lead to problems during model development. Missing data can be treacherous because it is difficult to identify the problem. You can’t predict when missing data are problematic because sometimes your results are affected and sometimes they are not. Also, it is not always obvious when missing data will cause a problem. Each question or variable may only have a small number of missing responses, but in combination, the missing data could be numerous. Only thorough analysis on your missing data can determine whether missing data are problematic.

Missing data can cause serious problems. First, most statistical procedures automatically eliminate cases with missing data. This means that in the end, you may not have enough data to perform the analysis. For example, you could not run a factor analysis on just a few cases. Second, the analysis might run but the results may not be statistically significant because of the small amount of input data. Third, your results may be misleading if the cases you analyse are not a random sample of all cases.

The prediction of missing values can be deal by using following methods they ate as follows

- Remove rows with missing values – the removal of rows will works well if the values are missing randomly and if we don't lose too much of the dataset after doing so.
- Build another predictive model to predict the missing values - This is a simple techniques which is used here.
- Use a model that can incorporate missing data - Like a random forest, or any tree-based method.

The genomics data, for example, in any missing data situations following imputation (fill-in) methods may-use

- Impute the missing value by the mean of the respective column.
- Imputation by linear/logistic regression.
- Imputation by KNN etc.,

RELATED WORK
There were some researchers worked on time series analysis data focusing in pre-processing with various strategies. Lippi et.al. Reviewed existing methods on short term traffic flow by Probabilistic graphical models and introduced two new support vector regression models for accuracy prediction of missing data [3]. By using comparison and SVR using the methodologies of time series model like Box and Jerkins, ARMA, ARIMA, SARIMA, KALAMAN filter. The intent of the paper is to predict the traffic conditions in a transport network on its past behavior. The advantage of the models used here is high accuracy in prediction of the traffic congestion. Finally it proves that kalamaban filter is the best model for the forecasting the short term traffic flow.

The objective of the paper in [1] is to detect the usual traffic congestions in spatio-temporal traffic data. Here in detection of the traffic congestions the existing methods are not effective so, proposed a new methodology to identify the traffic congestions which overcome the disadvantages of the existing system by using a STC Tree methodology. The approach of modelling the congestion propagation network using DBN is presented. The existing system uses the correctness and completeness frequent sub-tree algorithm and Apriori algorithm for frequent item set. The
proposed algorithms are validated by experiments on a large real-time travel time data in an urban road network. The proposed method for mining the congestion propagation patterns is a combination of STC and DBN. Each method individually has its own advantages and limitations. We use the STC algorithm to discover the frequent causal trees which is applicable for the large network. DBN is used to calculate the distribution of congestion propagation from the small sub-structures which were discovered by STC.

Shi, Zhu et.al.[2] focused on the filling the gaps in the missing data and classify the fault type in power grid by integrated data pre-processing framework [DPF] on apache spark using the methodology of missing data prediction by applying time series model for missing data prediction. And performing linear regression. To find the fault type classification we use logistic regression and support vector machine. The algorithms which are used here are Train BTSM, optimized linear regression for the Data Pre-processing Framework and for to implement on spark we use the missing data prediction algorithm and Fault type classification.

Chaouch proposed strategy for functional time series forecasting household-level electricity demand which represents a key factor to assure the balance supply/demand in the LV network [4]. A novel methodology able to improve short term functional time series forecasts has been introduced. Based on improvement of the FWK approach which is performed by classification, assigning and applying FWK. To get a reasonable assumptions new approach of forecasting was implemented i.e., “functional Time Series”. The intra-day individual load curve shape is mainly affected by the consumption behaviour of the customer and there is no evidence to identify a common pattern between days we used an unsupervised classification method to find similar segments. The proposed method will improved by using some daily exogenous functional random variables, like internal/external daily temperature and sunshine curves.

Ruiz et.al, proposed methodologies for time series analysis, based on micro blogging activity by using correlation and aggregation in financial time series [5]. This is of Features in the first group measure the overall activity in the micro-blogging platform, such as number of posts, number of re-posts, and so on. Features in the second group measure properties of an induced interaction graph, for instance, the number of connected components, statistics on the degree distribution, and other graph-based properties. Here they presented framework to extract messages from twitter and gave the detailed experimental results by measuring the correlation of the stock market events with these features, using Twitter which is a data source. Their results show that the most correlated features are the number of connected components and the number of nodes of the interaction graph. The correlation is stronger with the traded volume than with the price of the stock. Lastly, they were used simulation to show that these features are useful in order to improve a trading strategy in the stock market.

Li et.al, proposed a new fast algorithm for the mining of coevolving time series which mine the time series data effectively [6]. This can be done by following process first we need to do the mining with missing data next applying the Natural motion Stitching later we use a technique called Parallelization. They applied their algorithm to solve practical problems including occlusions in motion capture, and generating natural human motions by stitching low-effort motions. They also proposed a parallel learning algorithm for LDS to fully utilize the power of multicore/multiprocessors, which will serve as corner stone of many applications and algorithms for time series. The algorithm shows the good result but took co-evolving sequences. The future direction of this paper is to mining of the large co-evolving sequences.

Asif et.al, developed a framework to overcome the problem of missing data in ITS(Intelligent transport system) proposed a model for data imputation in large road networks that is Low-dimensional models in presence of missing historical and neighbouring data [7]. Traditional methods require the availability of sufficient historical data and assume that missing data occurs at isolated instances. Due to highly erratic reporting patterns of sensors, these assumptions are usually not valid in practical road networks. As a consequence, these methods fail to deal with the problem of missing data in large interconnected urban settings. To overcome these limitations, they proposed few methods that can perform data imputation by constructing low-dimensional models of large and
diverse networks in presence of missing data. That can be implemented by Fixed point continuation which uses SVD after applying the Missing data imputation by CP Decomposition and Missing data imputation using BPCA. The imputation accuracy of the proposed methods can potentially be improved by developing kernel versions of the methods. The future direction may be to predict the accuracy of the data driven traffic forecasting.

Ma et.al, focused on the memory-based collaborative filtering problems by means of two crucial factors: similarity computation between users or items and missing data prediction algorithms [8]. Here they proposed an effective missing data prediction EMDP algorithm for collaborative filtering by comparing metrics using Mean Absolute error (MSE). First, they used the enhanced Pearson Correlation Coefficient (PCC) algorithm by adding one parameter which overcomes the potential decrease of accuracy when computing the similarity of users or items. Second, they proposed an effective missing data prediction algorithm, where the information of both users and items is taken into account.

In [9], methodologies were proposed to predict the missing data by optimized SVM and Refined SVM and proposed a framework called OR_MLF which is machine learning framework. That is measured using the Mean Absolute Error. The experimental results show that the predictors based on the proposed framework achieve lower mean square error than traditional ones. The framework used here OR_MLF would be a good candidate to predict the missing data in power grid system where this framework OR_MLF with RSVM obtained the lowest MSE 0.0021.

CONCLUSION

It is an interesting problem to predict the number of citations a paper would receive. Though forecasting and prediction is not very accurate, it would be good if we could achieve higher percentage of accuracy. Another interesting idea would be to cluster the papers based on the time series. We reviewed some of the important works of multi variable time series analysis and Also we had discuss all the imputation methods for finding the missing values and also present the advantage and disadvantage of the different imputation methods for finding the missing value from the dataset in the field of data mining

REFERENCES

[1] H. Nguyen, W. Liu, F. Chen, “Discovering congestion propagation patterns in spatio-temporal traffic data”, IEEE Transactions on Big Data PP (99) (2016) 1–1.
[3] M. Lippi, M. Bertini, P. Frasconi, “Short-term traffic flow forecasting: An experimental comparison of time-series analysis and supervised learning”, IEEE Transactions on Intelligent Transportation Systems 14 (2) (2013) 871–882.
[2] W. Shi, Y. Zhu, T. Huang, G. Sheng, Y. Lian, G. Wang, Y. Chen, “An integrated data preprocessing framework based on apache spark for fault diagnosis of power grid equipment”, Journal of Signal Processing Systems 82 (2016) 1–16.
[4] M. Chaouch, “Clustering-based improvement of nonparametric functional time series forecasting: Application to intra-day household-level load curves”, IEEE Transactions on Smart Grid 5 (1) (2014).
[5] E. J. Ruiz, V. Hristidis, C. Castillo, A. Gionis, A. Jaimes, “Correlating financial time series with micro-blogging activity”, in: Proceedings of the Fifth ACM International Conference on Web Search and Data Mining, WSDM ’12, ACM, 2012, pp. 513–522.
[6] L. Li, C. Faloutsos, “Fast algorithms for time series mining”, 2010 IEEE 26th International Conference on Data Engineering Workshops (ICDEW 2010) 00 (2010) 341–344.
[7] M. Asif, N. Mitrovic, L. Garg, J. Dauwels, P. Jaillet, “Low-dimensional models for missing data imputation in road networks”, in: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2013, pp.
[8] H. Ma, I. King, M. R. Lyu, “Effective missing data prediction for collaborative filtering”, in: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and
Development in Information Retrieval, SIGIR '07, ACM, 2007, pp. 39–46.

[9] G. Wang, Y. Chen, “Improving power grid monitoring data quality: An efficient machine learning framework for missing data prediction”, in: 2015 IEEE 17th International Conference on High Performance Computing and Communications, IEEE, 2015, pp. 417–422.

[10] Weiwei Shi, Yongxin Zhu, Philip S. Yu, Jiawei Zhang, Tian Huang, Chang Wang, Yufeng Chen, “Effective Prediction of Missing Data on Apache Spark over Multivariable Time Series”. Transactions on Big Data, 2017.

[11] Yehuda Koren, “Matrix Factorization Techniques for Recommender Systems,” Published by the IEEE Computer Society, IEEE, pp. 42- 49, 2009.

[12] Huseyin Polat and Wenliang Du, “SVD-based Collaborative Filtering with Privacy,” ACM Symposium on Applied Computing SAC’, ACM, 2005.

[13] Manolis G. Vozalis and Konstantinos G. Margaritis, “A Recommender System using Principal Component Analysis,” Published in 11th Panhellenic Conference in Informatics, pp. 271-283, 2007.

[14] Jialu Liu Jiawei Han, “HINMF: a Matrix Factorization method for clustering in heterogeneous information networks”

[15] Weiwei Shi, Yongxin Zhu, Philip S. Yu, Tian Huang, Chang Wang, Yishu Mao, And Yufeng Chen, “Temporal Dynamic Matrix Factorization For Missing Data Prediction In Large Scale Coevolving Time Series”, IEEE Access, 2016.

[16] Makridakis, Spyros, “A Survey of Time Series”

[17] Soheila Mehrmolaei and Mohammad Reza Keyvanpour, “A Brief Survey on Event Prediction Methods in Time Series”, Springer International Publishing Switzerland 2015,

[18] Dr. P. Radha and R. Divya, “Survey on the Multiple Time Series Data with Data Mining Techniques”, International Journal of Computational Intelligence Research ISSN 0973-1873 Volume 13, Number 4 (2017), Pp. 615-620

[19] Weiwei Shi, Yongxin Zhu, Philip S. Yu, Mengyun Liu, Guoxing Wang, Zhiliang Qian,Yong Lian “Incomplete Electrocardiogram Time Series Prediction”, IEEE, 2016.

[20] Helmut Lütkepohl, “New Introduction to Multiple Time Series Analysis” Springer, 2015.

[21] Z. Asha Farhath, B. Arputhamary, Dr. L. Arockiam, “A Survey on ARIMA Forecasting Using Time Series Model”, IJCSMC, Vol. 5, 2016, Pg.104 – 109.

[22] Dheeraj Kumar Bokde, Sheetal Girase, Debajyoti Mukhopadhyay, “Role of Matrix Factorization Model in Collaborative Filtering Algorithm: A Survey”, International Journal of Advance Foundation and Research in Computer (IJAFRC) Volume 1, 2014.