Bidirectional Imputation of Sensor-based Time Series Data

Yuhui Wang
School of Electrical Engineering and Computer science
Washington State University
Pullman, WA, USA
mark166.wang@wsu.edu

Diane J. Cook
School of Electrical Engineering and Computer science
Washington State University
Pullman, WA, USA
djcook@wsu.edu

ABSTRACT

Although missing data is a frequent phenomenon in collected data, few efforts have focused on imputing a whole block of missing continuous time series data. With the purpose of helping wearable technology overcome data limitations resulting from battery constraints, we propose a time series ensemble model (TEM), which can estimate a sequence of missing time series data based on existing data. When TEM is used, the accuracy of activity recognition yields improved performance compared with random guess of activity. Therefore, we conclude that TEM can potentially improve the benefit of collecting continuous data by imputing the data missing between the collection periods.

KEYWORDS

Data Imputation, Time Series, Smart Watch, Sensor Data, Ensemble Method

Introduction

The use of wearable technology is expeditiously increasing. Data collected from wearable devices can help to recognize current activities and ultimately better understand human behavior and its influences. As an example, the long-term goal of our research is to learn the relationships between genetics, sensor-observed behavior, and health conditions. To find these connection, we need large amounts of continuous, longitudinal sensor data. We automatically label the collected data with activity categories and extract relevant features such as times spent on selected activities and changes in activity patterns over time. Smart watches offer a ubiquitous and ecologically valid method to collect continuous data. However, battery constraints place limitations on data collection. When wearable technology collects continuous sensor and location data, the battery drains quite fast, usually within 6 hours. Because round-the-clock data collection cannot be accomplished with a single device, we rely on data imputation to estimate the missing data between the collection periods. This presents a way to reduce difficulties in collecting continuous data via smartwatches during a long time period.

Missing data imputation has been widely investigated in the field [2,3,5,9,11,16,18]. Investigated approaches include EM [6], multiple imputation [1,17], kernel methods [15], and matrix factorization [4]. Additionally, Sovilj et al. introduced a data imputation method based both on a Gaussian Mixture Model and an Extreme Learning Machine [16]. The Gaussian Mixture Model generates a model to handle missing data based on a Gaussian distribution, and the Extreme Learning Machine generates a multiple imputation strategy for the final estimation. In recent years, Recurrent Neural Networks have demonstrated state-of-the-art performance in many applications with sequential data [12]. Che et al. have developed a deep neural network GRU-D which is based on a Gated Recurrent Unit (GRU).[2] This model can capture long-term dependences in time series and utilize the missing pattern for better predicting the result. However, few study works on entire period of data missing. Their studies majority focus on partial of the data missing in one query.

In this paper, we construct a time series ensemble model (TEM) for imputing smartwatch data. In contrast with previous approaches, we seek to generate an entire sequence of missing data rather than a single missing value. Based on the dependence between each sensor, we build a dependent tuning algorithm (DTA). DTA can help tune the final value of each sensor reading based on other sensor output. We validate our approach based on smart watch data collected for multiple subjects.

Method

We propose an ensemble method for time series data imputation. The five base estimators we use are linear regression [14], K-nearest neighbors regression [8], SVM (polynomial kernel and Gaussian kernel) [19] and polynomial regression [14]. These classifiers increase the diversity of the model, thus we hypothesize that combining these classifiers can improve model accuracy. We combine the base classifiers with voting. We further improve the method using Adaboost [10] to weight past errors from the model and further improve imputation accuracy. Figure 1 displays the architecture of our ensemble method for smart watch data imputation.

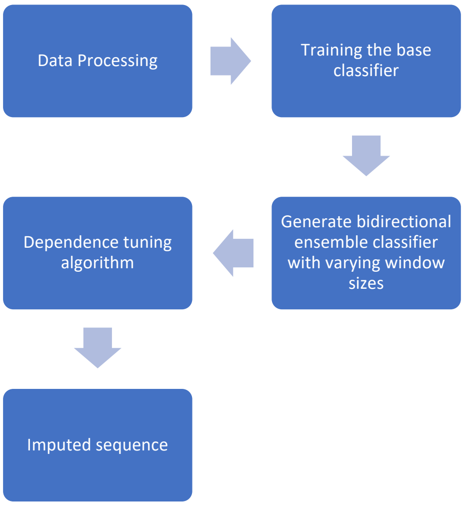
Data Processing
Training the base classifier
Generate bidirectional ensemble classifier with varying window sizes
Dependence tuning algorithm
Imputed sequence

Figure 1. TEM software architecture.

To define data points for analysis, we move multiple windows through the sensor data and extract features from each window. One set of windows moves forward in time, another set moves backward. By considering data in both directions we obtain a better contextual perspective of the missing data. When a block of data is missing from the time series, features from windows before the block and after the block are used to impute the missing block, as shown in Figure 2. Furthermore, we use multiple window sizes to generate training data with different context sizes. In our experiments we specifically consider windows of length 5, 10, 15, 20, 25, 30 and 60 sensor readings. In the forward direction, a predictor learns a mapping from window-based features to a value for the time step immediately following the window. In the backward direction, the predictor maps window-based features to a value for the time step immediately preceding the window.

Figure 2 illustrates the feature extraction and prediction process for a time series with length 15. In this figure, the blue entries represent a window of time series values for feature extraction and the red entries represent the predicted value. As shown in the middle of the figure, a forward-moving window extracts features from data points at times 1 through 5 and predicts the value at time 6, while the backward-moving window extracts features from data points 11 through 15 and predicts the value at time 11. As shown on the right, the forward window advances by one time step to repeat the process, the backward window moves backward one time step and repeats the process. The prediction process halts when the forward window (or backward window) reaches the end of the missing data block. In our experiments, we predict a maximum of 5 minutes of missing data, beyond which the integrity of imputed data is endangered.

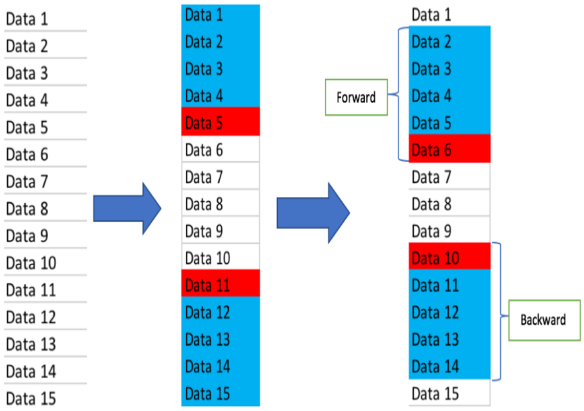

Figure 2. The processing of generating data points from raw time series data. Here we assume a window size of 5 sensor readings.

Because we utilize multiple window sizes (7 in our case) and 2 directions, we train 14 separate ensemble regressors on available data. Our training data and testing data consists of every continuous sequence of ground truth data with size (2 * maximum window size) + (length of imputed data block). For example, if we wish to impute 5 minutes (300 seconds) of data then we use data from sequences of length 2*60 + 300 = 420 sensor readings, or seconds of data, to create data points for training and testing the 14 learning algorithms. Finally, we train a meta-linear regressor to map output from the 14 base regressors onto a predicted value. The predicted value is then used as part of the input window to predict a value that is 2 time steps away from ground truth data. We repeat the process as need to predict the entire block of missing time series data. The result is our time-series ensemble method (TEM) for time series sensor data imputation.

After building TEM, we can optionally refine the model using joint prediction[7,13]. Joint inference allows us to consider the dependence between each imputed sensor type in the final predictor. We hypothesize that including this information can improve the accuracy of time series data imputation. To perform joint inference, we first use TEM to independently predict the value for each of the n sensors, s = {s_1, .., s_n}. In a second pass, when we predict the value of sensor 1, we use the predicted values for each of the other sensors, s_2 through s_n, to expand the input feature vector. The regressors use the original feature vector for sensor 1 together with the predicted values of s_2.. s_n to output the predicted value for s_1. The same process is applied in a second learning pass for the other sensors as well. We refer to this joint prediction algorithm as TEM-DTA, which contrasts with the original single-pass TEM algorithm.

Experiment Setup

We hypothesize that the bi-directional, ensemble method contained in TEM will accurately impute large blocks of missing sensor-based time series data. If this hypothesis is validated, the resulting technique will allow researchers to obtain longitudinal wearable data for human behavior analysis even when the actual data collection is sparse. To validate our hypothesis, we train and test TEM using actual smartwatch sensor data.

Our dataset is collected by 4 participants who each wore Apple Watch 2 continuously for multiple days. Collected sensor data include yaw, roll, pitch, rotation rate (X,Y,Z), acceleration (X,Y,Z), speed, latitude, longitude and altitude. These data are collected once per second. All participants are graduate students (1 female and 3 males). The participants wear the watch in their daily lives, but the period of data collection varies among each participant. The longest data collection period is 2 months, and the shortest one is 2 weeks. In total, we have 185,974 seconds of collected sensor data. Participants collected data continuously throughout the day but took off the watch to charge it at night. As a result, we utilized each separate day of data for creating data points.

In addition to collecting raw sensor data, participants also provide activity labels. In this way we can determine not only the accuracy of imputed data with respect to observed ground truth, but we can also quantify the impact of TEM on applications that utilize the information such as activity recognition. For this dataset, participants provided activity labels every 5 minutes while wearing the watch. Activities are labeled as work, exercise, relax, eat, walk or other. Because the value range of each sensor type is different, we normalize the data from 0 to 1 before we train the model. Since the sensor data include negative numbers, we apply Equation 1 for this normalization.

$$\overline{X_n} = \frac{X_n - X_{\min}}{X_{\max} - X_{\min}} \tag{1}$$

In our experiment, we report results based on three-fold cross validation. We evaluate TEM and TEM-DTA using multiple performance measures. The measurements we use are mean absolute error (MAE) and root mean squared error (RMSE). In addition to evaluating the accuracy of imputed data, we also evaluate the impact of imputed data on activity recognition performance. Specifically, we compare activity recognition performance based on data imputed by TEM, by TEM-DTA, and by TEM using only a single direction (only forward or only backward windows). We utilize an activity recognition model that was trained on data provided by a larger group of 20 participants using a random forest with 100 trees and entropy feature-selection criteria.

Results

In order to evaluate TEM, we construct two imputation models for two of the five base regressors: one is KNN and another is SVM with Gaussian kernel (GK). Of the five, KNN has a rich history of performing well on this type of data and the SVM performed best as a stand-alone regressor in our experiments. The two models use single directional (forward and backward) and bidirectional moving windows.

We compare the performance of TEM and TEM-DTA with a selection of other data imputation methods on different data block sizes. Figure 4 shows the MAE and RMSE for the tested classifiers. “Forward only” and “backward only” regressors utilize TEM in just one direction, while “Average” represents the average of all the voting regressors and all window sizes for forward-only and backward-only choices. In the two graphs, we can see the one-direction classifiers perform worse compared with bidirectional algorithms. TEM and TEM-DTA dramatically outperform the other algorithms. TEM-DTA also exhibits a slight increase in performance over TEM alone, providing evidence that capturing the relationship between predicted variables improves imputation accuracy.

For all regression strategies, error increases monotonically with the size of the continuous imputed data block. This is because the missing data rate, or proportion of missing data (imputed data) to total data (imputed data and windows on either side) grows with the data block size. In Figure. 4, we see that even though the blocking period increases, the performance of TEM and TEM-DTA decreases slower than that of the other algorithms.

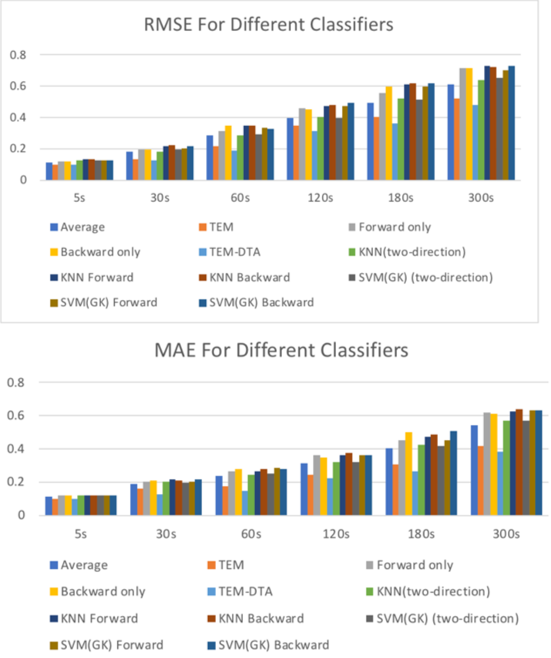

Figure 4. MAE (top) and RMSE (bottom) for selected regression strategies.

After determining the RMSE and MAE measures for alternative classifiers, we use pre-trained activity recognition model to test the accuracy of the data we imputed. The pre-trained Random Forest model yields 85.19% recognition accuracy for the four-person dataset. We compare multiple data imputation methods, and apply 3-fold cross validation to measure accuracy of a model trained with the resulting data.

Figure 5 graphs the results of all the imputation methods. We can see that, with 5s missing data, the performance of all data imputation methods is similar. As the missing data rate increases with larger imputed data blocks, the performance TEM and TEM-DTA drops more slowly than for the other methods. This provides evidence that the bi-directional ensemble regression is stronger than individual regressors, particularly as larger blocks of continuous data are imputed.

Furthermore, the performance of all methods that rely on only one direction drops nearly to the level of random guess as the missing data period reaches 300s, while the accuracy of the bidirectional methods still remain above 30%. This indicates that the bidirectional scanning provides important multiple perspectives on the data, yielding more accurate predicted values.

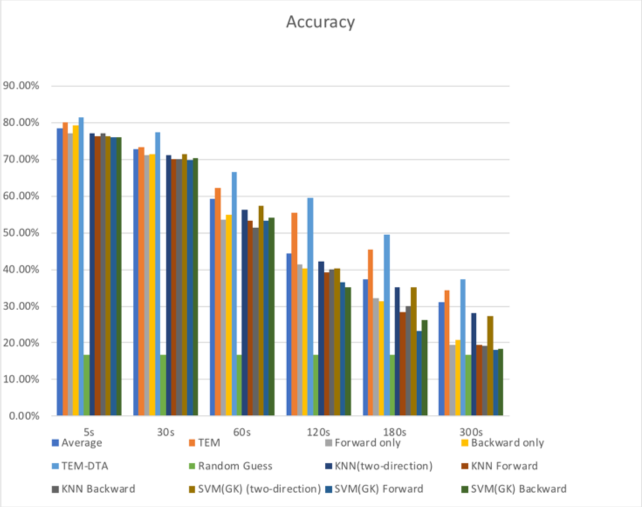

Figure 5. Activity recognition accuracy for selected data imputation methods.

Conclusion

In this paper, we propose a new method for solving the problem of missing time series data. We hypothesized that an ensemble which considers a diversity of regressors, context directions, and window sizes can provide accurate data imputation even for continuous blocks of time series data. The results of our experiments show that our method, TEM-DTA, yields reasonable performance even when imputing as much as 300s missing data. When imputing 5s missing data, TEM-DTA yields 81.34%, which lowers recognition accuracy less than 4% from the original baseline performance.

Reference

[1] Melissa J. Azur, Elizabeth A. Stuart, Constantine Frangakis, and Philip J. Leaf. 2011. Multiple imputation by chained equations: What is it and how does it work? *Int. J. Methods Psychiatr. Res.* (2011). DOI:https://doi.org/10.1002/mpr.329

[2] Zhengping Che, Sanjay Purushotham, Kyunghyun Cho, David Sontag, and Yan Liu. 2018. Recurrent Neural Networks for Multivariate Time Series with Missing Values. *Sci. Rep.* (2018). DOI:https://doi.org/10.1038/s41598-018-24271-9

[3] Xiaobo Chen, Cheng Chen, Yingfeng Cai, Hai Wang, and Qiaolin Ye. 2018. Kernel sparse representation with hybrid regularization for on-road traffic sensor data imputation. *Sensors (Switzerland)* (2018). DOI:https://doi.org/10.3390/s18092884

[4] B. J. Cowling, G. Freeman, J. Y. Wong, P. Wu, Q. Liao, E. H. Lau, J. T. Wu, R. Fielding, and G. M. Leung. 2013. Spectral Regularization Algorithms for Learning Large Incomplete Matrices. *Euro Surveill. Bull. Eur. sur les Mal. Transm. = Eur. Commun. Dis. Bull.* (2013). DOI:https://doi.org/10.1016/j.surg.2006.10.010.Use

[5] A. Rogier T. Donders, Geert J.M.G. van der Heijden, Theo Stijnen, and Karel G.M. Moons. 2006. Review: A gentle introduction to imputation of missing values. *J. Clin. Epidemiol.* (2006). DOI:https://doi.org/10.1016/j.jclinepi.2006.01.014

[6] Pedro J. García-Laencina, José Luis Sancho-Gómez, and Aníbal R. Figueiras-Vidal. 2010. Pattern classification with missing data: A review. *Neural Comput. Appl.* (2010). DOI:https://doi.org/10.1007/s00521-009-0295-6

[7] Shima Ghassem Pour and Federico Girosi. 2016. Joint prediction of chronic conditions onset: Comparing multivariate probits with multiclass support vector machines. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics).* DOI:https://doi.org/10.1007/978-3-319-33395-3_13

[8] Trevor Hastie and Robert Tibshirani. 1996. Discriminant adaptive nearest neighbor classification. *IEEE Trans. Pattern Anal. Mach. Intell.* (1996). DOI:https://doi.org/10.1109/34.506411

[9] W. L. Junger and A. Ponce de Leon. 2015. Imputation of missing data in time series for air pollutants. *Atmos. Environ.* (2015). DOI:https://doi.org/10.1016/j.atmosenv.2014.11.049

[10] Hui Liu, Hong Qi Tian, Yan Fei Li, and Lei Zhang. 2015. Comparison of four Adaboost algorithm based artificial neural networks in wind speed predictions. *Energy Convers. Manag.* (2015). DOI:https://doi.org/10.1016/j.enconman.2014.12.053

[11] Q. Ma, Y. Gu, F.-F. Li, and G. Yu. 2016. Order-sensitive missing value imputation technology for multi-source sensory data. *Ruan Jian Xue Bao/Journal Softw.* (2016). DOI:https://doi.org/10.13328/j.cnki.jos.005045

[12] Fabiola Mancini, Andrea Savarino, Maria Losardo, Antonio Cassone, and Alessandra Ciervo. 2009. Neueal Mechine Translation BY JOINTLY LEARNING TO ALIGN AND TRANSLATE. *Microbes Infect.* (2009). DOI:https://doi.org/10.1016/j.micinf.2008.12.015

[13] Andreas Peldszus and Manfred Stede. 2015. Joint prediction in MST-style discourse parsing for argumentation mining. DOI:https://doi.org/10.18653/v1/d15-1110

[14] Sunil J Rao. 2009. Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis. *J. Am. Stat. Assoc.* (2009). DOI:https://doi.org/10.1198/jasa.2003.s263

[15] K. Rehfeld, N. Marwan, J. Heitzig, and J. Kurths. 2011. Comparison of correlation analysis techniques for irregularly sampled time series. *Nonlinear Process. Geophys.* (2011). DOI:https://doi.org/10.5194/npg-18-389-2011

[16] Dušan Sovilj, Emil Eirola, Yoan Miche, Kaj Mikael Björk, Rui Nian, Anton Akusok, and Amaury Lendasse. 2016. Extreme learning machine for missing data using multiple imputations. *Neurocomputing* (2016). DOI:https://doi.org/10.1016/j.neucom.2015.03.108

[17] Ian R. White, Patrick Royston, and Angela M. Wood. 2011. Multiple imputation using chained equations: Issues and guidance for practice. *Stat. Med.* (2011). DOI:https://doi.org/10.1002/sim.4067

[18] D. Alexandra Williams, Benjamin Nelsen, Candace Berrett, Gustavious P. Williams, and Todd K. Moon. 2018. A comparison of data imputation methods using Bayesian compressive sensing and Empirical Mode Decomposition for environmental temperature data. *Environ. Model. Softw.* (2018). DOI:https://doi.org/10.1016/j.envsoft.2018.01.012

[19] Huanhuan Yuan, Guijun Yang, Changchun Li, Yanjie Wang, Jiangang Liu, Haiyang Yu, Haikuan Feng, Bo Xu, Xiaoqing Zhao, and Xiaodong Yang. 2017. Retrieving soybean leaf area index from unmanned aerial vehicle hyperspectral remote sensing: Analysis of RF, ANN, and SVM regression models. *Remote Sens.* (2017). DOI:https://doi.org/10.3390/rs9040309