Missing Data Estimation Framework of Air Pressure System using Gaussian Processes Regression

Eunseo Oh
Industrial Engineering
Kumoh National Institute of Technology
Gumi, South Korea
chss014@kumoh.ac.kr

Hyunsoo Lee
Industrial Engineering
Kumoh National Institute of Technology
Gumi, South Korea
hsl@kumoh.ac.kr

Abstract— **As companies with high quality and reliability dominate the market in the automotive market, a failure analysis framework is considered an essential framework for guarantying reliability in modern manufacturing and production processes. In automotive industry, it is important to identify failure causes of a vehicle. However, a set of collected manufacturing data may have missing values in several attributes in automotive manufacturing data. Due to these missing values, data imbalance and a lack of defect-related data and imbalance might occur in its learning process. Then, it may give result to inaccurate failure analysis. In order to overcome these issues, a framework to handle missing values using Gaussian Processes Regression (GPR) is proposed. The proposed framework makes it possible to estimate unbalanced sample data and missing data are interpolated. In order to show the effectiveness of purposed framework, GPR-based missing data interpolation framework is analyzed and tested by Air Pressure System (APS) failure of Scania trucks data. Then, the results are compared with other existing methods.**

Keywords—Air Pressure System, Missing Data Estimation, Data, Gaussian Process Regression (GPR), Data Mining

I. Introduction

In order to analyze and classify the cause of product's failure in several manufacturing equipment processes, a number of studies have been conducting actively to solve problems by applying machine learning or deep learning methods. In particular, it is important to identify the cause of the failure and malfunction of vehicles and to predict their character data before the failure caused by the Air Pressure System (APS) in automotive industry. Therefore, it is necessary to distinguish between vehicle defects/ malfunctions caused by APS and other than APS. These days, a number of machine learning methods have been applied to these failure cause analyses, and they show decent performances when relevant data are enough and well distributed. However, attribute data obtained in the actual plant process is disproportionately distributed, and a model trained with this data might give result to bad classification performance. Many existing studies have been presented as methods of handling missing values to overcome this data imbalance problem. There have been methods for handling missing values with median [1] and for handling missing values of process data using the Generative Adversary Network (GAN) [2].

This research study proposes a new and effective framework to classify pass or fail after handling missing values through Gaussian Processes Regression (GPR) [4] with high flexibility using a real APS data.

Section 2 presents a handling mechanism model for missing values using GPR. Section 3 shows its effectiveness with the experimental results of the proposed framework and comparisons with other classification models using APS data.

II. Missing data handling framework using Gaussian Processes Regression

In general, a real data obtained in manufacturing equipment processes have issues such as lack of data and data imbalance including many missing values. Such data, including imbalances or missing data, make it difficult to analyze the data accurately. In this study, in order to estimate the missing values, the missing values labeled with "NaN-Not a Number" are estimated using the average value and re-interpolated using GPR.

First, the missing value is replaced by the average value of the attribute data in case of the existence of a missing value using equation (1).

$$\frac{\sum_{i=1}^{n} a_i}{n} \tag{1}$$

n is the number of remaining values except for the missing value in the property in which the missing value exists. Then, the missing values are estimated using GPR by selecting a set of instances. The selected data generate a distribution then a sample from the estimated distribution replaces the initially estimated median value.

(2), (3), (4) and (5) summarize the general GPR model.

$$\mathrm{y} = \mathrm{f} + \epsilon \tag{2}$$

where, $\epsilon \sim \mathrm{N}\left(0, \sigma_y^2 \mathrm{I}\right)$

$$\begin{bmatrix} \boldsymbol{Y} \\ \boldsymbol{f}_* \end{bmatrix} \sim \boldsymbol{N}\left(\boldsymbol{0}, \begin{bmatrix} \boldsymbol{k}(\boldsymbol{X},\boldsymbol{X}) + \boldsymbol{\sigma}_y^2\mathbf{I} & \boldsymbol{k}(\boldsymbol{X},\boldsymbol{X}_*) \\ \boldsymbol{k}(\boldsymbol{X}_*,\boldsymbol{X}) & \boldsymbol{k}(\boldsymbol{X}_*,\boldsymbol{X}_*) \end{bmatrix}\right) \tag{3}$$

$$\mathrm{m}(X_*) = \mathrm{k}(X_*, \mathrm{X})^T (\mathrm{k}(\mathrm{X},\mathrm{X}) + \sigma_y^2 \mathrm{I})^{-1} Y \tag{4}$$

$$\mathrm{V}[f_*] = \mathrm{k}(X_*, X_*) - \mathrm{k}(X_*, \mathrm{X})^T (\mathrm{k}(\mathrm{X},\mathrm{X}) + \sigma_y^2 \mathrm{I})^{-1} k(X, X_*) \tag{5}$$

In (2), ϵ is a noise parameter with a mean of 0 and a variance of σ_y^2 subsequent to the estimated Gaussian distribution, and I is an Identity matrix.

The pre-distribution of intermediate value f_* for converting the input value X to the observed value Y is modeled using (3).

Using the updated Gaussian process, the driven Maximum Likelihood Estimator (MLE) and its variance are estimated using (4) and (5), respectively.

Finally, the missing value is estimated by predicting a new estimate Y_* through the (4) and (5). The newly generated data is learned as training data for classification of pass and fail.

III. Experiment and result

One of the important issues in classification through machine learning is how to handle missing data. In this study, missing values are estimated using GPR.

In order to show the effectiveness of the proposed method, a real data - APS Failure at Scania Trucks Data Set [3] is used. The training data consists of 170 attributes and 60,000 instance vectors. The identification of each individual is divided into 59,000 failure cause data due to APS and 1,000 failure cause data instead of APS. The test data is divided into 955 failure cause data due to APS and 45 failure cause data instead of APS. Several attributes in each data have missing values or NaNs. Table 1 shows the number of missing values in APS data .

TABLE I. Missing value issues in APS data

Categories	***Total size***	***data item with missing values***	***Percentages***
Number of attributes	171	171	100%
Data set	60,000	59,998	99.99%

In order to estimate these missing terms, the proposed method is applied. The interpolated data is trained and tested with a deep neural network (DNN) as shown in Fig. 1. The used DNN model has one input layer, five hidden layers and one output layer. The objective of the DNN model is classify the pass and fail of APS data.

Fig.1 shows a deep neural network structure.

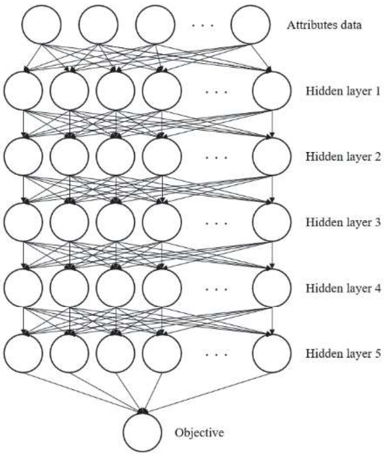

Fig. 1. Deep neural network structure

Fig. 2.

Table 2 shows the classification errors of three models: 1) the model without sampling, 2) the model replaced only with the mean value, and 3) the proposed model.

TABLE II. Comparison of classification error

Compared methods	***Error***
Proposed Model (GPR-based missing value estimation)	0.0749
Data ignorance with missing values	0.2061
Mean Imputation	0.1612

Each method uses a DNN model with the same number of hidden layers (the number of hidden layer =5) for classifying the data. The each weight vectors are determined with each generated data set. Then, it is experimentally proved that the missing value handling using the proposed GPR-based method is effective for the following classification processes.

IV. Conclusion and further studies

The problem of failure classification for improving the quality and reliability of vehicles has been studied, and the classification performance have been improved using various machine learning methods.

In order to improve the classification performance of vehicle failures with missing values, this study uses GPR-based interpolation method. Initially, missing values of an attribute are replaced with the mean values, and the data are estimated using GPR.

The validity of the generated data was verified by a subsequent classification model with the relevant classification errors. It is confirmed that the proposed model is effective with a real industrial data and the comparisons with several existing methods. Further studies are expected to provide a framework for unbalanced data-based problems in multiple classes and more effective analytics.

Acknowledgment

This research was supported by The Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, S. Korea (grant number: NRF-2018R1D1A3B07047113) and by Priority Research Centers Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (2018R1A6A1A03024003).

References

[1] C. Gondek, D. Hafner and O. R. Sampson, "Prediction of Failures in the Air Pressure System of Scania Trucks using a Random Forest and Feature Engineering," International Symposium on Intelligent Data Analysis, Springer, Cham, pp. 398-402, September, 2016.

[2] H. Kim and H. Lee, "Fault Detect and Classification Framework for Semiconductor Manufacturing Processes using Missing Data Estimation and Generative Adversary Network," Journal of Korean Institute of Intelligent Systems, vol. 28, pp. 393-400, 2018.

[3] T. Lindgren and J. Biteus, “UCI Machine Learning Repository,” [Online].Available: https://archive.ics.uci.edu/ml/datasets/APS+Failure+at+Scania+Trucks.

[4] C. K. I. Williams and C. E. Rasmussen, “Gaussian Processes for Regression,” Advances in Neural Processing Systems, vol. 8, pp. 514-520, 1996.