Neurocomputing 174 (2016) 220–231

Contents lists available at ScienceDirect

Neurocomputing

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

Extreme learning machine for missing data using multiple imputations

Dušan Sovilj [a,d,*], Emil Eirola [a], Yoan Miche [b], Kaj-Mikael Björk [a], Rui Nian [c], Anton Akusok [d], Amaury Lendasse [a,d]

[a] *Arcada University of Applied Sciences, 00550 Helsinki, Finland*
[b] *Nokia Solutions and Networks Group, 02022 Espoo, Finland*
[c] *Ocean University of China, 266003 Qingdao, China*
[d] *The University of Iowa, Iowa City, IA 52242-1527, USA*

ARTICLE INFO

Article history:
Received 22 August 2014
Received in revised form
15 March 2015
Accepted 17 March 2015
Available online 10 August 2015

Keywords:
Extreme Learning Machine
Missing data
Multiple imputation
Gaussian mixture model
Mixture of Gaussians
Conditional distribution

ABSTRACT

In the paper, we examine the general regression problem under the missing data scenario. In order to provide reliable estimates for the regression function (approximation), a novel methodology based on Gaussian Mixture Model and Extreme Learning Machine is developed. Gaussian Mixture Model is used to model the data distribution which is adapted to handle missing values, while Extreme Learning Machine enables to devise a *multiple imputation* strategy for final estimation. With multiple imputation and ensemble approach over many Extreme Learning Machines, final estimation is improved over the mean imputation performed only once to complete the data. The proposed methodology has longer running times compared to simple methods, but the overall increase in accuracy justifies this trade-off.

1. Introduction

Recurring problem in many scientific domains is the accurate prediction or forecast for unknown and/or future instances. This issue is addressed by assuming that there exist the underlying mechanism that generates the available data, and then building a model that provides good enough approximation for that same mechanism. Finally, any kind of inference is based on the constructed model assuming all the necessary information is taken into account. The task of making predictions, for example, daily temperature or retail sales for some specific time period, is considered a *regression* problem or estimation of the regression function. Another issue becoming more prevalent in machine learning domain is related to the missing data in databases encountered in many research areas [1–4]. This issue has huge impact on both the learning algorithms and the subsequent inference procedures. If this issue is not treated correctly, any kind of inference results in severely biased and inaccurate estimates.

In the paper, we are interested with regression problems of the form:

$$y_i = f(\mathbf{x}_i) + \epsilon_i \qquad (1)$$

in the presence of missing data where $(\mathbf{X};\mathbf{Y}) = \{(\mathbf{x}_i;\mathbf{y}_i)\}_{i=1}^{N}$ are data samples with $\mathbf{x}_i$ consisting of d explanatory features or variables, y_i the target variable and ϵ_i the noise term. The usual assumption behind the noise term is that it follows a Gaussian distribution with zero mean and known variance $\epsilon \sim \mathcal{N}(0, \sigma^2)$. The regression problem is to find a model $\mathcal{M}$ that is a close approximation to the true underlying function f. The case explored in this paper is when samples or observations $\mathbf{X}$ contain unobserved (unknown) variables, that is, the values are missing for certain observation and features. Values could be missing for a variety of reasons depending on the source of the data, including measurement error, device malfunction, operator failure, and many others. On the other hand, many modelling methods assume that data contain a fixed number of samples lying in a fixed feature space. Presence of missing values prevents these methods to be applied directly to the data. Simple *ad hoc* solutions to incomplete data include completely removing samples containing unobserved values, mean imputation with the mean computed on available values, replacement from correlated variables, substitution based on prior information (such as regional codes) and others. In order to provide more reliable inference after the modelling stage, a suitable strategy must be employed.

[*] Corresponding author at: The University of Iowa, Iowa City, IA 52242-1527, USA.
E-mail addresses: dusan-sovilj@uiowa.edu (D. Sovilj), emil.eirola@arcada.fi (E. Eirola), yoan.miche@nokia.com (Y. Miche), kaj-mikeal.bjork@arcada.fi (K.-M. Björk), anton-akusok@uiowa.edu (A. Akusok), amaury-lendasse@uiowa.edu (A. Lendasse).

http://dx.doi.org/10.1016/j.neucom.2015.03.108
0925-2312/© 2015 Elsevier B.V. All rights reserved.

Besides simple ad hoc procedures, there are several paradigms for dealing with missing data used in conjunction with machine learning methods [5] and these include:

- *Conditional mean imputation* approach which is optimal in terms of minimising the mean squared error of the imputed values, but suffers from biased statistics of the data. For instance, estimates of variance or distance are negatively biased.
- *Random draw imputation* that is more appropriate for generating a representative instance of a fully imputed data set. However, the imputations can be highly variable with respect to any single value to be accurate.
- *Multiple imputation*: This setup draws several representative imputations of the data, analyses each set separately, and combines the results to form an overall estimate with uncertainty taken into account [6]. This approach can result in unbiased and accurate estimates after a sufficiently high number of draws, but it is not always straightforward to determine the posterior distribution to draw from [7]. In the context of machine learning, repeating the analysis several times is impractical as training and analysing a sophisticated model tends to be computationally expensive.

The conceptually simplest approach to dealing with incomplete data is to fill in the missing values before commencing any further analysis. Many methods have been suggested for imputation with the intent to appropriately conform to the distribution of the data. These include imputation by nearest neighbours [8], or the improved incomplete-case k-NN imputation [9]. An alternative approach is to study the input density indirectly through conditional distributions by fully conditional specification [10]. However, the uncertainty of the imputed values is often not explicitly modeled in most imputation methods, and hence ignored in the further analysis, potentially leading to biased results.

Having an appropriate model to take into consideration missing data has several advantages. First, with any kind of imputation, many learning algorithms can be directly applied to imputed data, such as neural networks, Gaussian processes and density estimation methods. Second, having a specific model designed to tackle missing values allows to take into consideration the variability of imputed values, and thus, the variance of the final estimation the practitioner is interested about.

Finite mixture models are a powerful modelling tool with a wide array of applications. Of considerate importance is the Gaussian Mixture Model (GMM), also known as Mixture of Gaussians, which has been studied extensively to describe distributions of a data set. This model provides a suitable estimation of the underlying data density distribution as GMM is a universal approximator [11]. This enables GMM to model any kind of continuous densities to arbitrary precision, and has been employed for a variety of problems in vision [12,13], language identification [14], speech [15,16] and image [17,18] processing. The parameters of GMM are obtained via maximum likelihood (ML) estimation by the Expectation–Maximisation (EM) algorithm [19]. EM algorithm is a general purpose algorithm for finding the ML solution with latent variables or incomplete data and does not require any derivatives of the likelihood function. GMM has been extended to accommodate missing values in data sets [20,21] which has seen some resurgence in recent years [22–24].

In this paper, we are considering regression estimation in the presence of missing data. First, mixture of Gaussians is applied to original data with missing values. Second, a large number of imputations is performed, that is, a multiple imputation approach is adopted. After all newly formed data sets are available, a suitable regression model is build. As the number of draws can be large and the data sets can often contain huge number of samples, a fast (in terms of training speed) and accurate model should be used. The choice is on Extreme Learning Machine (ELM) as it satisfies both criteria. In the case of difficult data, where substantial number of imputed data sets is required, ELM acts a good model as fast computational models are more viable than the alternative gradient-based neural networks or kernel methods.

Gaussian Mixture Model has been used to train neural networks in the presence of missing data [25] with the average gradient computed for the relevant parameters by using conditional distribution for the missing values. The method is designed to handle training of networks with back-propagation and is not applicable to other machine learning methods. Extreme Learning Machine has also been adapted to handle missing values [26,27] with both approaches estimating distances between samples that are subsequently used for the RBF kernel in the hidden layer. One advantage of that approach is circumventing estimation of all the missing values and focusing only on providing required information for the methods based on distances, such as Support Vector Machines or k-nearest neighbours. However, the method only returns expected pairwise distances that are then employed by the ELM for regression. The downside is that other activations functions have to be ignored, and the imputation is done once by the conditional mean. Although conditional mean imputation provides improved results over simple ad hod solutions, it neglects the variability introduced by the underlying Gaussian Mixture Model.

The rest of the paper is organised as follows: Section 2 explains the overall approach in more detail focusing on the main points in the methodology. Two main components of the approach, namely Mixture of Gaussians for missing data and Extreme Learning Machine are explained in Sections 3 and 4 respectively. Section 5 showcases the results between two types of imputation – conditional mean and multiple imputation, combined with two different modelling strategies. Finally, summarising remarks are given in Section 6.

2. Methodology

The overall approach consist of four consecutive stages:

1. Fitting the Gaussian Mixture Model on a data set with missing values.
2. Generating new data sets via multiple imputation based on the Gaussian Mixture Model from the first stage.
3. Building Extreme Learning Machine for each generated data set in the second stage.
4. Combining all the Extreme Learning Machines to provide final estimates.

2.1. Gaussian Mixture Model fitting

In the first stage, a Gaussian Mixture Model Γ is fitted to the data with missing values. Since the data contains missing values, straightforward application of the EM algorithm is not possible and certain adjustments are necessary for both E and M-steps. In the E-step, conditional expectations with respect to known values in samples are used to obtain means and covariances for the missing values. In the M-step, the conditional mean fills the missing parts (per sample imputation) in order to compute GMM component means. The covariance matrices for each component are similarly adjusted taking into account covariances for the missing parts. The details required to carry out these corrections are explained in Section 3.1.