ESKİŞEHİR TEKNİK ÜNİVERSİTESİ BİLİM VE TEKNOLOJİ DERGİSİ B- TEORİK BİLİMLER

Eskişehir Technical University Journal of Science and Technology B- Theoretical Sciences

2020, 8(2), pp. 336 - 343, DOI: 10.20290/estubtdb.747821

DEEP LEARNING-BASED APPROACH FOR MISSING DATA IMPUTATION

Pınar CİHAN [1, *]

[1] Computer Engineering, Çorlu Engineering Faculty, Tekirdağ Namık Kemal University, Tekirdağ, Turkey

ABSTRACT

The missing values in the datasets are a problem that will decrease the machine learning performance. New methods are recommended every day to overcome this problem. The methods of statistical, machine learning, evolutionary and deep learning are among these methods. Although deep learning methods is one of the popular subjects of today, there are limited studies in the missing data imputation. Several deep learning techniques have been used to handling missing data, one of them is the autoencoder and its denoising and stacked variants. In this study, the missing value in three different real-world datasets was estimated by using denoising autoencoder (DAE), k-nearest neighbor (kNN) and multivariate imputation by chained equations (MICE) methods. The estimation success of the methods was compared according to the root mean square error (RMSE) criterion. It was observed that the DAE method was more successful than other statistical methods in estimating the missing values for large datasets.

Keywords: Deep learning, Autoencoder, Denoising autoencoder, Missing data

1. INTRODUCTION

Deep learning is a member of the machine learning family, and its popularity has increased considerably in recent years. Deep learning is a highly effective method in the training of very large and complex probabilistic models and the ability to train these systems. Thanks to the latest developments in deep learning, technology has achieved great results in many areas [1]. Especially, deep neural networks provide very good results when compared to other traditional approaches. It autoencoder (AE), which is a derivative of artificial neural networks, consists of three layers; input layer, hidden layer and output layer. A simple AE structure is shown in Figure 1.

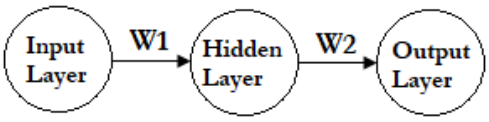

Figure 1. The structure of AE

Because the AE uses the same input dataset as the output of the network, it does not require labels containing the class information in the dataset. Therefore, it is an unsupervised machine learning method. The structure of AE consists by encoder and decoder parts. A sample AE neural architecture with an attribute number of 7 and 4 neurons in the hidden layer is as in Figure 2. The AE has a feed forward structure and it can have multiple hidden layers. The main purpose in the AE model is to try to recreate the value of x_i in the input layer in the output layer x_i'. The weights (w, w') are continuously updated by means of the back propagation algorithm in order to make these two values come close together. In the AE, the size of the output layer is equal to the size of the input layer. This is the most fundamental difference that distinguishes the autoencoder from the traditional neural network. While the encoder part reduces the multidimensional input information to small dimensions, the decoder part returns the data back to its original structure [2, 3].

*Corresponding Author: pkaya@nku.edu.tr
Received: 04. 06.2020 Published: 31.08.2020

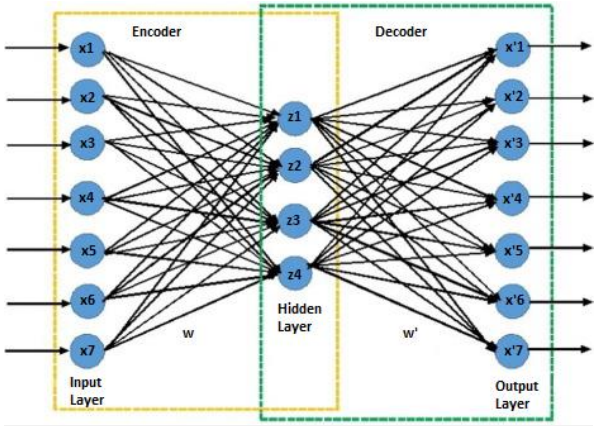

Figure 2. Sample of AE neural architecture [4]

In the literature, the AE are used in many applications such as noise reduction [5], data compression [6, 7], abnormal pattern detection [8], outlier detection [9], data visualization [10], speech recognition [11]. The stacked autoencoder (SAE) and the denoising autoencoder (DAE) are AEs version. Although these methods have been developed for other purposes, they have recently become successful methods of handling missing data [12].

There are several ways to handle missing data (Figure 3): Case deletion, imputation of missing values, model-based procedures and machine learning methods. Although all of these methods have the advantage and disadvantage, recently, these methods are frequently used in the literature for handling missing values. These approaches are usually based on statistics and machine learning, and the missing values are replaced with plausible values. In statistical methods, missing values are imputed with values such as mean/median of missing values without creating a predictive model. In machine learning based techniques, a predictive model is created using the complete data and the missing values are imputed with the result of this model.

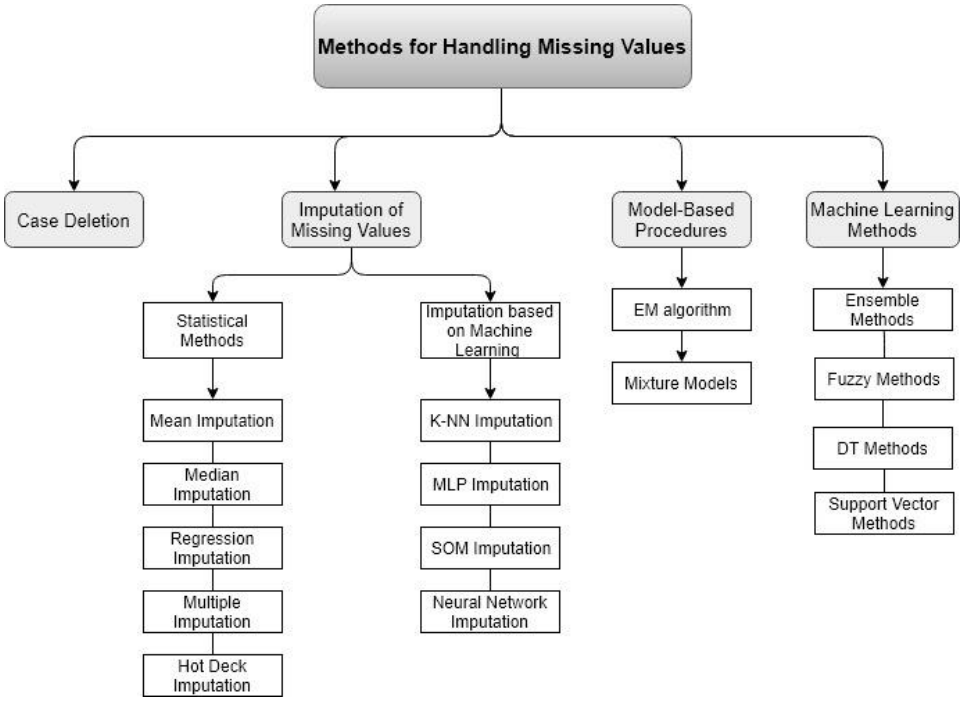

Figure 3. Methods for handle missing values (adapted from [13])

Nowadays, even though deep learning techniques still popular [1], the imputation of the missing values with these methods does not receive enough attention. When missing values in very large datasets are imputed by non-model based methods such as kNN, SVM, SOM, it causes serious disadvantages in terms of computational cost.

Recently, imputation of missing values using by the neural network based methods show the up and upward momentum. One of the most suitable architectures for this purpose is AEs. This type of network learns a representation of the data in the input layer and tries to reproduce it on the output layer. By doing this, the model can learn from missing data and generate new reasonable values for those that are missing. In this study, the missing value imputation studies by using the AEs will be examined and the studies will be compared to how the AE performs according to the other missing value imputation methods. In addition, missing values in three different datasets will be estimated by using DAE, KNN and MICE methods and the prediction success of these methods will be compared according to RMSE.

2. RELATED WORK

Duan et al. [14] traffic data was used in the study and the dataset was taken from Caltrans Performance Measurement System (PeMS). Missing values at rates 1 - 90% were simulated on dataset. The missing values in the dataset have been imputed by the stacked denoising autoencoder (SDAE) and the artificial neural network (ANN). According to the RMSE and mean absolute error (MAE) criteria, the SDAE method was found to be more successful than the ANN method. As a result of the study, the authors reported that the deep learning method was a good method for imputing the traffic dataset.

Duan et al. [15] expanding their previous work and compare the proposed SDAE method with different imputation methods. In the study, missing data at rates of 5 - 50% were generated on traffic data. They have compared the proposed SDAE method with the history model, the integrated moving average (ARIMA) and the BP neural network model. As a result of the study, they concluded that SDAE method once again performed better to other methods.

Gondara and Wang [16] compared DAE with the MICE method. They used synthetic and real-world datasets which taken from UC Irvine Machine Learning Repository (UCI). 60% and 80% missing values are simulated under the Missing completely at random (MCAR) and missing not at random (MNAR) mechanism. As a result of the study it has been reported that according to the RMSE and accuracy criteria, the DAE method perform better result in both datasets.

Gondara and Wang [17] work is very similar to the previous study. There are two different missing data scenarios in this study. One, all features are set to missing values that is uniform synthetic production. Other, half of the features are set to missing values that is random synthetic generation. The results reported that the DAE approach was more successful than MICE in all the uniform scenarios and in 7 cases for the random scenario.

Beaulieu-Jones and Moore [18] stated that the missing data is very common in the field of electronic health records and this situation creates difficulties. They used ALS EHR dataset taken from pooled resource open-access ALS clinical trials (PRO-ACT). In the study, missing data at rates of 10 – 50% were simulated on the ALS EHR dataset. AE approach was compared with mean, median, SoftImpute, kNN and SVD methods. As a result of the study it has been reported that the AE method was more successful than other methods.

Zhao et al. [19] proposed a fast cluster-based SAE method (SAE-FC). In study, four different datasets used which are taken from UCI. The aim of the study was to increase the clustering accuracy and the proposed method was compared with expectation maximization imputation (EMI), fuzzy c-means (FCM), FIMUS and DMI methods. As a result of the study, it has been reported that the proposed

method is more successful than all other methods, and the use of SAE has nearly doubled the success of clustering in some scenarios.

Shao et al. [20] proposed autoencoder based approach to overcome image classification with missing modality. Four datasets used taken from BUAA and Oulu-CASIA databases. Proposed methods compared with Transfer Subspace Learning (TSL), Low-rank Transfer Subspace Learning (LTSL), Robust Domain Adaptation with Low Rank Reconstruction (RDALR) and Geodesic Flow Kernel (GFK) methods. As a result of the study, it has been reported that new SAE approach was more successful in all scenarios than the other methods, mean accuracy was between 73.47% and 89.83% and the remaining algorithms did not exceed 70%.

Tran et al. [21] proposed cascaded residual autoencoder (CRA) to impute missing modalities. They used GRSS data fusion contest dataset (GRSS), the RGB-D object dataset (RGB-D), multi-PIE (MTPIE), and the hyperspectral face dataset from Hong Kong Polytechnic University. In the study, CRA compared with singular value threshold (SVT), SoftImpute, OptSpace, GA, DAE, SDAE and other AEs. As a result of the study, it has been reported that CRA with optimization or convolutional CRA achieves the best performance among all methods in all datasets.

Malek et al. [22] proposed pixel-based AE and patch-based AE methods to recover missing data in multispectral images due to the presence of clouds. Pixel-based uses a normal SAE that receives as input the images pixels, patch-based uses in SAE for each patch and results are merge by weighted average. In the study two datasets was used. First dataset acquired by FORMOSAT-2 satellite and second dataset acquired by the French satellite SPOT-5. These AE methods compare with Basis Pursuit, the orthogonal matching pursuit (OMP) and genetic algorithms (GAs). As a result of the study, it has been reported that proposed pixel-AE and patch-AE methods show good results in reconstructing the missing areas and can significantly outperform methods.

Ning et al. [23] compared the SDAE with the Weighted k-nearest neighbours data filling algorithm based on grey correlation analysis (GBWkNN) and the mutual k-nearest neighbours Imputation (MkNNI) methods. SDAE method shows better performance than other methods by RMSE and run time.

3. MATERIAL AND METHODS

In the study, three different real-life datasets [24] were used for comparisons of imputation models. Table 1 shows the properties of datasets. MCAR mechanism was used for creating missingness. Approximately 35% missing values are simulated under the MCAR mechanism and were performed five times for each dataset. These missing values in datasets were estimated by DAE [12], KNN, and MICE methods. Finally, performance of imputation models was evaluated using RMSE. The RMSE is calculated between the actual and the estimated values as follows:

$$\sqrt{\frac{1}{M}\sum_{i=1}^{M}(E_i - A_i)^2}$$

Where M is the number of missing values, E_i is an imputed (estimated) value and A_i is an actual value of the i^{th} missing value.

Table 1. Used datasets' properties for model evaluation

Dataset	Samples	Attributes
DNA	3186	180
Wine	4898	12
Shuttle	58000	9

DAE-based imputation

AEs are composed of input, hidden and output layers. AEs consists of two parts which are encoder and decoder. Encoder, goes from the input layer to the output of the hidden layer. It converts the input vector X into a hidden representation Y through a nonlinear transformation function. This is done by minimizing the loss function between the input X and the output of network Y. RMSE is commonly used as the loss function. Decoder, goes from the hidden layer to the end of the output layer. It maps the hidden embedding to the reconstruction of input X. DAE is a variant of AE often used for handling of missing values. DAE designed to recover the noiseless, original input through deep networks when a noisy input is given. The main difference of DAE and AE is the training phase. In training phase noise is added such as gaussian noise or salt-and-pepper to the input X to obtain X′ [25,26]. In this study, to handling the missing data with the DAE method in the "R" programming language, the *h2o.deeplearning* method in package *h2o* was used [27].

kNN-based imputation

kNN imputation technique based on a single imputation method. While estimate missing values with this method, firstly it searches for its nearest k neighbor of the instance. Then, imputing missing value using a weighted mean of the neighbors. Since it uses similarity distance for prediction, it takes time to find the nearest instance for large datasets. However, prediction performance is quite good compared to methods such as the mean imputation method [28, 29]. In this study, to handling the missing data with the kNN method in the "R" programming language, the *kNN* method in package *VIM* was used [30].

MICE-based imputation

Multiple imputation using chained equations technique based on a multiple imputation method. To estimating missing values, it uses regression methods. Missing values estimated n times and the final imputed dataset is calculated by the average of the n imputed datasets [31]. In this study, to handling the missing data with the MICE method in the "R" programming language, the *mice* method in package *MICE* was used [32].

4. RESULTS

After simulating approximately 35% missing values five times on three datasets used in the study, these missing values were estimated by DAE, kNN and MICE methods. Imputation models compared using RMSE results calculated per attribute on the test set. The RMSE results obtained are given in the box-plot graph (Figure 4).

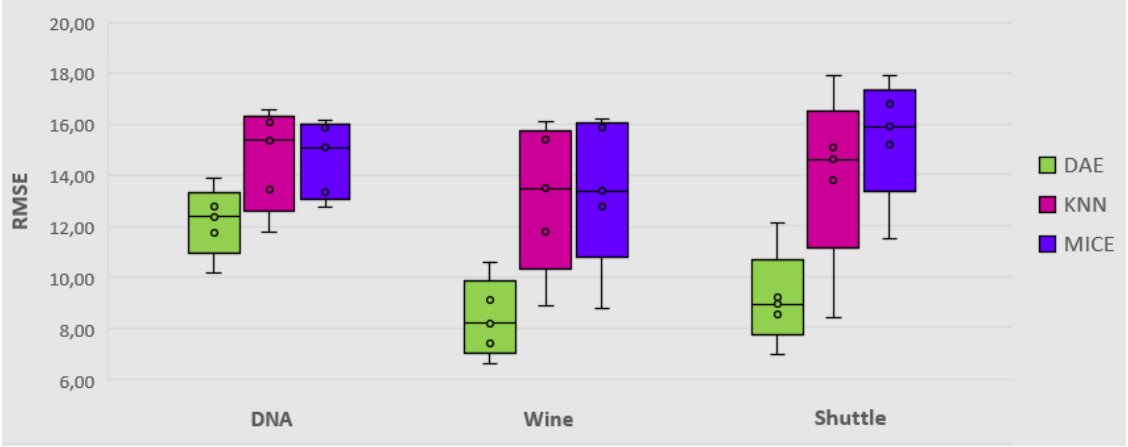

Figure 4. Comparison of imputation methods performance

In all datasets, the DAE method has the lowest average RMSE value. It is desirable that the RMSE value is low. Because, the lowest RMSE value is to show us the method estimation ability is high. In other words, the prediction value is close to the actual value. Deep architecture based models are known to be more successful in large datasets (high dimension, large sample), so one reason for the DAE method to be more successful than other methods is may that the datasets are large.

5. CONCLUSION

Real-world data is not always clean. Data sets often contain missing values. In order to make an analysis of datasets, the completion of missing values with close to actual values directly affects the analysis success. In this study, DAE performance, which is one of the deep learning methods that have gained popularity, has been compared with the statistical methods frequently used in the literature (KNN and MICE) to replace missing values. Also, studies examined which are used of autoencoder and its variants to handling the missing values in the datasets. It was observed that deep learning-based imputation methods showed outperforms to statistical methods. The AE and its variants are seen as very powerful and promising methods to handling the missing data. As a result of experimental study, DAE method, which is based on deep learning, was more successful than other methods in completing the missing values in the three large datasets used in the study.

REFERENCES

[1] Şeker A, Diri B, Balık HH. Derin Öğrenme Yöntemleri ve Uygulamaları Hakkında Bir İnceleme. Gazi Mühendislik Bilimleri Dergisi 2017; 3:47-64.

[2] Ballard DH. Modular Learning in Neural Networks. In: AAAI, 1987; pp 279-284.

[3] Qiu YL, Zheng H, Gavaert O. A deep learning framework for imputing missing values in genomic data. bioRxiv:406066 2018.

[4] Ahmed H, Wong M, Nandi A. Intelligent condition monitoring method for bearing faults from highly compressed measurements using sparse over-complete. features Mechanical Systems and Signal Processing 2018; 99:459-477.

[5] Ishii T, Komiyama H, Shinozaki T, Horiuchi Y, Kuroiwa S. Reverberant speech recognition based on denoising autoencoder. In: Interspeech 2013; pp 3512-3516.

[6] Del Testa D, Rossi M. Lightweight lossy compression of biometric patterns via denoising autoencoders. IEEE Signal Processing Letters 2015; 22:2304-2308.

[7] Tan CC, Eswaran C. Using autoencoders for mammogram compression. Journal of medical systems 2011; 35:49-58.

[8] Sakurada M, Yairi T. Anomaly detection using autoencoders with nonlinear dimensionality reduction. In: Proceedings of the MLSDA 2014 2nd Workshop on Machine Learning for Sensory Data Analysis 2014; p 4.

[9] Chen J, Sathe S, Aggarwal C, Turaga D. Outlier detection with autoencoder ensembles. In: Proceedings of the 2017 SIAM International Conference on Data Mining 2017; pp 90-98.

[10] Hinton GE, Salakhutdinov RR. Reducing the dimensionality of data with neural networks science 313:504-507.

[11] Lu X, Tsao Y, Matsuda S, Hori C. Speech enhancement based on deep denoising autoencoder. In: Interspeech 2013; pp 436-440.

[12] Vincent P, Larochelle H, Bengio Y, Manzagol P-A. Extracting and composing robust features with denoising autoencoders. In: Proceedings of the 25th international conference on Machine learning 2008; pp 1096-1103.

[13] García-Laencina PJ, Sancho-Gómez J-L, Figueiras-Vidal AR. Pattern classification with missing data: a review. Neural Computing and Applications 2010; 19:263-282.

[14] Duan Y, Lv Y, Kang W, Zhao Y. A deep learning based approach for traffic data imputation. In: Intelligent Transportation Systems (ITSC), 2014 IEEE 17th International Conference on 2014; IEEE, pp 912-917.

[15] Duan Y, Lv Y, Liu Y-L, Wang F-Y. An efficient realization of deep learning for traffic data imputation. Transportation research part C: emerging technologies 2016; 72:168-181.

[16] Gondara L, Wang K. Recovering loss to followup information using denoising autoencoders. In: 2017 IEEE International Conference on Big Data (Big Data) 2017; pp 1936-1945.

[17] Gondara L, Wang K Mida. Multiple imputation using denoising autoencoders. In: Pacific-Asia Conference on Knowledge Discovery and Data Mining 2018; pp 260-272.

[18] Beaulieu-Jones BK, Moore JH. Missing data imputation in the electronic health record using deeply learned autoencoders. In: PACIFIC SYMPOSIUM ON BIOCOMPUTING 2017.;World Scientific, pp 207-218.

[19] Zhao L, Chen Z, Yang Z, Hu Y, Obaidat MS. Local similarity imputation based on fast clustering for incomplete data in cyber-physical systems. IEEE Systems Journal 2018; 12:1610-1620.

[20] Shao M, Ding Z, Fu Y. Sparse low-rank fusion based deep features for missing modality face recognition. In: Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on 2015; pp 1-6.

[21] Tran L, Liu X, Zhou J, Jin R. Missing Modalities Imputation via Cascaded Residual Autoencoder. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2017; pp 1405-1414.

[22] Malek S, Melgani F, Bazi Y, Alajlan N. Reconstructing Cloud-Contaminated Multispectral Images With Contextualized Autoencoder Neural Networks IEEE Transactions on Geoscience and Remote Sensing 2018; 56:2270-2282.

[23] Ning X, Xu Y, Gao X, Li Y. Missing data of quality inspection imputation algorithm base on stacked denoising autoencoder. In: Big Data Analysis (ICBDA), IEEE 2nd International Conference on 2017 IEEE 2017; pp 84-88.

[24] Leisch F, Dimitriadou E. Machine Learning Benchmark Problems. R Package, mlbench, 2010.

[25] Vincent P, Larochelle H, Bengio Y, Manzagol PA. Extracting and composing robust features with denoising autoencoders. In Proceedings of the 25th international conference on Machine learning 2008; pp 1096-1103.

[26] Vincent P, Larochelle H, Lajoie I, Bengio Y, Manzagol PA, Bottou L. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. Journal of Machine Learning Research 2010; 11: 3371–3408.

[27] Gondara L, Wang K. Mida: Multiple imputation using denoising autoencoders. In Pacific-Asia Conference on Knowledge Discovery and Data Mining 2018; pp 260-272, Springer, Cham.

[28] Batista GE, Monard MC. An analysis of four missing data treatment methods for supervised learning. Applied artificial intelligence 2003; 17(5-6): 519-533.

[29] Hron K, Templ M, Filzmoser P. Imputation of missing values for compositional data using classical and robust methods. Computational Statistics & Data Analysis 2010; 54(12): 3095-3107.

[30] Templ M, Alfons A, Kowarik A, Prantner B. VIM: Visualization and Imputation of Missing Values. R package version 4.6.0, 2016, URL https://CRAN.R-project.org/package= VIM.

[31] White IR, Royston P, Wood AM. Multiple imputation using chained equations: issues and guidance for practice. Statistics in medicine 2011; 30(4): 377-399.

[32] Buuren SV, Groothuis-Oudshoorn K. mice: Multivariate imputation by chained equations in R. Journal of statistical software 2010; pp 1-68.