Novel Imputation Method Using Average Code from Autoencoders in Clinical Data

Edwar Macias, Javier Serrano, Jose Lopez Vicario, Antoni Morell
Wireless Information Networking (WIN) group
Universitat Autònoma de Barcelona (UAB)
08193 Bellaterra, Spain
{edwar.macias, javier.serrano, jose.vicario, antoni.morell}@uab.cat

***Abstract*—It is possible to improve the reconstruction of clinical data combining codes from autoencoders (AE). The extracted information can be used for enhancing existing imputation methods in this type of data. In the proposed approach, initially, encoder and decoder functions from trained autoencoder are extracted. Then, imputers equally spaced from normalized distribution of the variables generate codes that are combined in the average one that is finally used to reconstruct the original information. The proposed method is compared imputing by mean values of variables and using a single AE for reconstruction. The proposed approach has an outstanding performance recovering original information. It is even better with missing values in more than one variable. The error is at least 70% less than the other methods imputing one variable, and also the proposed approach is highly recommended with missing values in more than one variable.**

***Index Terms*—Imputation, deep learning, autoencoder, healthcare**

I. Introduction

In the era of deep learning (DL), it is possible to extract knowledge from clinical data even with the presence of missing values (MV). Those are generated due to lack of collections, errors in medical equipment readings, the omission of information by patients, or merging different sources of data that do not match in their timestamp. Thus, it is rare to have complete data sets, being necessary the application of imputation mechanisms that extract knowledge even with the presence of MVs.

In recent decades, the methods to impute MVs in the clinical domain have been based on simple imputations through constant values, imputing with previous measurements, or simply eliminating records with MVs [1]–[3]. More sophisticated imputation methods are based on multiple imputation [4], which combine several copies of the data, imputed by statistical measurements. However, these approaches do not consider complex relationships that data may have. On the other hand, DL has shown outstanding potential in exploiting hidden relationships in clinical data [5]–[7], being a promising alternative to impute MVs.

Using autoencoders (AE) as DL mechanism for data imputation has gained strength in recent years [8]–[11]. AEs use artificial neural networks (ANN) to replicate the input in the output layer, compressing the most relevant information in the hidden layers that allows the reconstruction with a reasonable error. The compressed representation is called codes. In the AEs approach, in the presence of MVs, three key stages are considered: i) replace MVs with constant or random values, ii) train an AE able to extract the most representative information and iii) track and replace the MVs with the reconstructed information from the AE. However, in the first stage, those values do not include most of the information from variables distribution. Therefore, the combination of imputed copies of data, with values that follow their distribution, could offer a better and robust reconstruction.

In this manuscript, a novel imputation mechanism based on the combination of generated codes using prior information from variable distribution is proposed. The main contributions of this manuscript are i) reconstruct MVs with a lower error using the average code and ii) offer insights into the effects of the imputation to different ratios of MVs and features.

In the next section, the MVs problem is explained. Section III presents the details of the proposed approach. In section IV, the approach is evaluated in two clinical data sets, and section V presents the remarks and conclusion of this work.

II. Problem Definition

Let $\mathbf{X} \in \mathbb{R}^{p,q}$ be a clinical dataset with p samples $\{\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(p)}\}$, where a sample $\boldsymbol{x}^{(i)}$ contains q features. Assuming the presence of MVs described by a p x q binary matrix $\mathbf{M}$ with '1' representing MVs and '0' observed values, see Fig. 1, the imputation goal is to replace the missing information with values that minimized bias in clinical studies, keeping as much information as possible. Thus, with a sample $\boldsymbol{x}^{(i)}$ imputation aims to find a function $h\left(\boldsymbol{x}^{(i)}\right)$ that replaces the MVs with the smallest error.

An AE allows the most representative structure of the data to be preserved in a smaller space through its encoder function, $f(\bullet)$. This function generates the so-called codes. Then, thanks to the decoder function, $g(\bullet)$, it is possible to use the codes to reconstruct the information with a relatively small error.

This work is supported by the Spanish Government under Project TEC2017-84321-C4-4-R co-funded with European Union ERDF funds and also by the Catalan Government under Project 2017 SGR 1670.

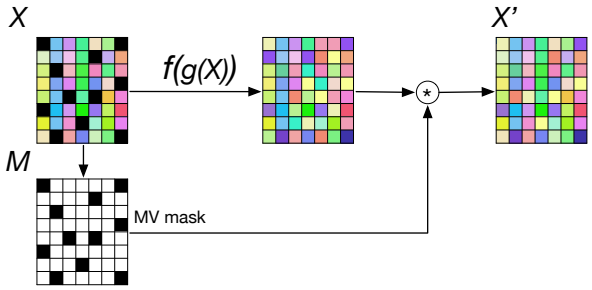

Fig. 1. Missing values reconstruction. Dark samples correspond to missing values.

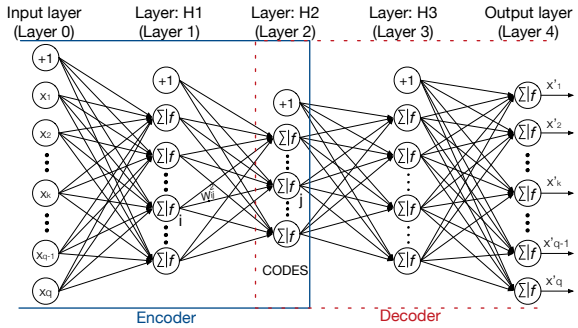

Fig. 3. Autoencoder structure.

III. PROPOSED METHOD

To exploit the reconstructive capacity of the average code, initially, the dataset $\mathbf{X}$ is split in training and test sets, $\mathbf{X}_{Train} \in \mathbb{R}^{m,q}$ and $\mathbf{X}_{Test} \in \mathbb{R}^{n,q}$, where $m+n=p$. An AE is trained with $\mathbf{X}_{Train}$, the encoder and decoder functions are extracted. Then, MVs in $\mathbf{X}_{Test}$ are imputed using values from the normalized distribution, $\mathcal{N}(0,1)$ from $\mathbf{X}_{Train}$. Finally, the average code is generated and the error calculated following the scheme in Fig. 2. Next, the stages of the framework are explained.

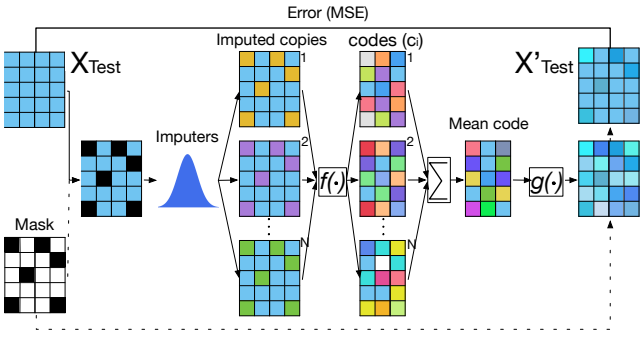

Fig. 2. Workflow for imputing and reconstruct the MVs, $\mathbf{X}_{Test}$ is the original data and $\mathbf{X'}_{Test}$ the reconstruction. $f(\bullet)$ and $g(\bullet)$ the encoder and decoder functions, respectively.

A. Knowledge extraction: Autoencoder

An AE is a type of ANN which replicates the input $\boldsymbol{x}^{(i)}$ to the output $\boldsymbol{x'}^{(i)}$ with the minimum possible error. It is composed of an encoder function, $f(\bullet)$, which represents the input in a latent space (codes), and a decoder, $g(\bullet)$, that takes this information and reconstructs the original input. Fig. 3 illustrates its components. The purpose of an AE is forcing the network to learn a data representation in a smaller space.

To train an ANN with L layers ($l = 0, \ldots, L-1$) and N_l neurons per layer, it is necessary to minimize iteratively a cost function that measures the error between the input and its reconstruction. The input is propagated through the connections, weights, of the units in the hidden layers. The output of the *i-th* neuron in the *l-th* layer, the so-called activation, is the linear combination of the outputs of the previous layer, taking the learned weights, and modified by a non-linear function, $y(\bullet)$. In other words, $a_j^l = y\left(\sum_{i=0}^{N_{l-1}} w_{i,j}^{l-1} a_i^{l-1}\right)$, where $w_{i,j}^{l-1}$ is the weight that connects the i-th activation at layer $l-1$ to the input of the j-th neuron at layer l. Note that $a_0^l = 1$ in all layer except the output layer to take into account the bias term. When the input is propagated to the output, the error is measured. In this work mean squared error (MSE) as cost function is used,

$$MSE = \frac{1}{N}\sum_{i=1}^{N}\left(\boldsymbol{x}^{(i)} - \boldsymbol{x'}^{(i)}\right)^2 \tag{1}$$

in this case the input $\boldsymbol{x}^{(i)}$ is used as target as well and $\boldsymbol{x'}^{(i)}$ are the predicted values.

The weights that present the smallest MSE are found through back-propagation algorithm [12], which uses the information of the gradient of the cost function to update the weights,

$$\mathbf{W}(t+1) = \mathbf{W}(t) - LR * \Delta MSE\left(\mathbf{W}(t)\right) \tag{2}$$

where $\mathbf{W}$ contains the ANN weights and learning rate (LR) controls how fast the error moves to a local minimum in the cost function.

To accelerate network learning, the LR can be adaptive. In this work, the adaptive momentum estimation (ADAM) [13] is used.

B. Imputation

The encoder and decoder functions, $f(\bullet)$ and $g(\bullet)$ in Fig. 2, are extracted from an AE trained with data without MVs from $\mathbf{X}_{Train}$. Next, MVs are generated in $\mathbf{X}_{Test}$, keeping the missing mask. Next, N copies of the $\mathbf{X}_{Test}$ are generated, and each one is imputed with N values equally spaced in the interval [-1, 1]. Then, each imputed dataset is passed by the encoder function and N codes, $c_i \in \mathbb{R}^{n,o}$ ($o < q$), are generated. Finally, the average code is computed (see Eq. 3) and passed by the decoder function to reconstruct the input and impute just the MVs.

$$\bar{\boldsymbol{c}} = \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{c}_i \tag{3}$$

IV. EXPERIMENTAL RESULTS

For testing the performance of the proposed approach, two public clinical datasets are used: Diagnostic Wisconsin breast cancer [14], with 30 continuous features and 569 samples. The second dataset contains information from patients with acute kidney injury (AKI) in intensive care units; data were extracted from the medical information mart for intensive care III (MIMIC III) [15]. It contains 28 continuous features and 4153 samples without MVs. Next, the proposed approach is evaluated in two scenarios. When MVs appear in one feature and, in a more realistic scenario, when they appear in several ones. In both cases, the reconstructive capacity is evaluated for different MV rates, from 10-50%.

Initially, the data from both datasets are randomly split into training, validation, and test set (70-15-15%). Later, two AEs with one hidden layer with 20 and 25 units (dimension of the codes) for cancer and AKI datasets, respectively, are trained. $LR = 0.005$ with ADAM optimizer and 'ReLU' as activation function was used. Then, the reconstructive capacity of the proposed approach is compared with two approaches. The classical approach [16], imputing with the mean value of the variables. The second approach uses a more sophisticated imputation mechanism used in the literature, [8], [17], [18], imputing directly with just one code form a trained AE. For this case, MVs are initially replaced with random values following the distribution of the variables. Then reconstruction is done following the scheme in Fig. 3. The main difference with the proposed approach is the initial replacement of MVs, authors in the literature replace them with random values from the distribution of the variables, while in ours, equally spaced values from the normalized distribution are used.

A. Reconstructing one variable

The influence of imputing MVs with constant values is evaluated measuring the MSE of equally spaced imputers in every feature. Thus, random MVs (10%) are generated. Then, they are imputed with 100 equally spaced values in the range [-1, 1]. Fig. 4 shows how the MSE evolves for the imputers in every feature in the test set of the cancer dataset. It can be appreciated how the minimum MSE does not correspond to the same imputer value for most of the features and also is not necessarily concentrated around the mean value of the variable.

Then, the average code was used to impute individual variables at different MV ratios. Fig. 5 shows its performance compared with the other methods. In overall, the MSE is smaller for most of the features. However, when the MV rate is higher than 20%, the proposed approach has an error of less than 70% compared with the other two imputation methods. In the AKI dataset, a similar effect happened.

B. Reconstructing more than one variable

Analyzing the impact of having MVs in multiple variables, the proposed approach was tested reconstructing MVs in several variables at the same time, starting with MVs in one variable, then in two, and so on. The reconstructive error on the cancer dataset can be appreciated in Fig. 6. It can be seen how the error rapidly increases when there are MVs in more than 18 variables with a missing ratio higher than 20%. However, the proposed approach retains a relatively constant error, and the number of variables does not significantly change the MSE compared to the other approaches. MSEs of the proposed approach are in the range 0.001-0.006, while the other ones have a minimum MSE of 0.005 for 10% of MVs in less than 30% of its variables.

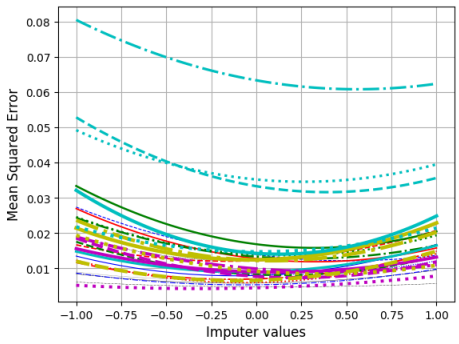

Fig. 4. Imputation using a constant value in cancer dataset. A total of 100 equally spaced values in the interval [-1, 1] are used. Each curve represent the error evolution in the features.

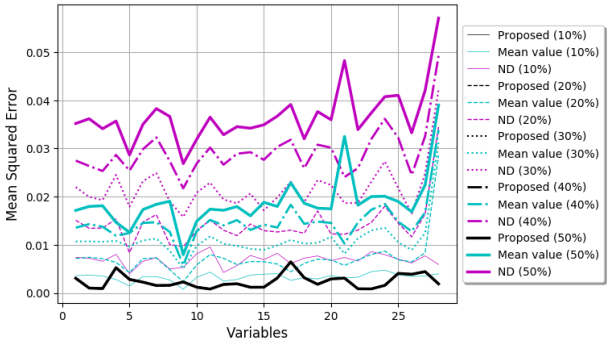

Fig. 5. Comparison of reconstructed values imputing by the mean codes (Proposed), the mean value of each variable and using AE to reconstruct the original value imputing MVs with random values in a normal distribution (ND) for the AKI dataset.

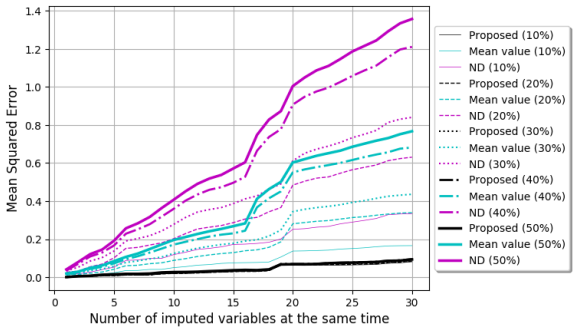

Fig. 6. Comparison in reconstruction for several variables at the same time using proposed approach, mean and random normal distributed values (ND) for the cancer dataset.

On the other hand, using the information in the initial experiment (Fig. 4), the values from the distribution with the smallest error for each variable are extracted as imputers and compared with the proposed approach. Then, they impute the variables and its performance is compared with the proposed approach using accumulative variables. In Fig. 7 and Fig. 8 it can be appreciated, for both datasets, how the approach has a smaller error, and it is significant with more than 14% of MVs in more than seven variables. Also, it can be examined how the error increases when the AE encodes more than 16 and 14 variables in AKI and cancer dataset, respectively. To the best of our knowledge, this is because there are values of variables that the AE is not able to reconstruct correctly from the code space. However, when using the average code, this behaviour is minimized by considering the values will be around the distribution of the variable.

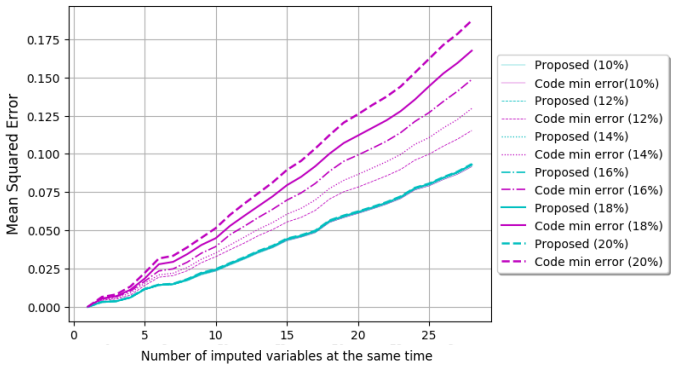

Fig. 7. Proposed method vs individual code with the minimum error in AKI dataset.

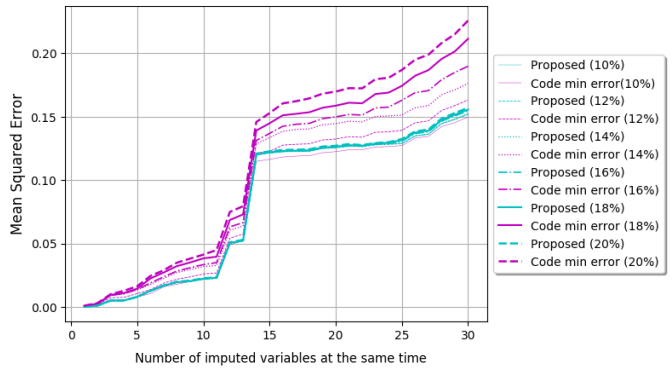

Fig. 8. Reconstruction by code with minimum error and proposed approach in cancer dataset.

V. CONCLUSION

In this work, a novel mechanism of imputation based on the average code using the normalized distribution of variables was proposed. It was compared with two imputations methods. The first one, commonly used in medicine by imputing with the mean value of the features; and the second one was imputing MVs with a trained AE. The proposed approach overcomes the other ones. It was analyzed how the average code performed better than the single ones generated from the imputation with values that a priori presented a minimum error in reconstruction. We recommend the use of the approach having more than 20% of MVs in a single variable. For the case of having MVs in more than one variable, the method has an outstanding performance having more than 14% of MVs in at least 30% of the variables.

REFERENCES

[1] Donald B Rubin, "Inference and missing data," *Biometrika*, vol. 63, no. 3, pp. 581–592, 1976.

[2] John W Graham, "Missing data analysis: Making it work in the real world," *Annual review of psychology*, vol. 60, pp. 549–576, 2009.

[3] Zhongheng Zhang, "Missing data imputation: focusing on single imputation," *Annals of translational medicine*, vol. 4, no. 1, 2016.

[4] Joseph L Schafer, "Multiple imputation: a primer," *Statistical methods in medical research*, vol. 8, no. 1, pp. 3–15, 1999.

[5] E Macias, A Morell, J Serrano, and JL Vicario, "Novel imputing method and deep learning techniques for early prediction of sepsis in intensive care units," in *2019 Computing in Cardiology Conference (CinC)*. IEEE, 2019, vol. 45, pp. 1–4.

[6] Travers Ching, Daniel S Himmelstein, Brett K Beaulieu-Jones, Alexandr A Kalinin, Brian T Do, Gregory P Way, Enrico Ferrero, Paul-Michael Agapow, Michael Zietz, Michael M Hoffman, et al., "Opportunities and obstacles for deep learning in biology and medicine," *Journal of The Royal Society Interface*, vol. 15, no. 141, pp. 20170387, 2018.

[7] E Macias, A Morell, J Serrano, and JL Vicario, "Knowledge extraction based on wavelets and dnn for classification of physiological signals: Arousals case," in *2018 Computing in Cardiology Conference (CinC)*. IEEE, 2018, vol. 45, pp. 1–4.

[8] Lovedeep Gondara and Ke Wang, "Mida: Multiple imputation using denoising autoencoders," in *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, 2018, pp. 260–272.

[9] Ramiro D Camino, Christian A Hammerschmidt, and Radu State, "Improving missing data imputation with deep generative models," *arXiv preprint arXiv:1902.10666*, 2019.

[10] Emimal Jabason, M Omair Ahmad, and MN S Swamy, "Missing structural and clinical features imputation for semi-supervised alzheimer's disease classification using stacked sparse autoencoder," in *2018 IEEE Biomedical Circuits and Systems Conference (BioCAS)*. IEEE, 2018, pp. 1–4.

[11] Guillem Boquet, Jose Lopez Vicario, Antoni Morell, and Javier Serrano, "Missing data in traffic estimation: A variational autoencoder imputation method," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 2882–2886.

[12] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams, "Learning representations by back-propagating errors," *nature*, vol. 323, no. 6088, pp. 533–536, 1986.

[13] Diederik P Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[14] Dheeru Dua and Casey Graff, "UCI machine learning repository," 2017.

[15] Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark, "Mimic-iii, a freely accessible critical care database," *Scientific data*, vol. 3, pp. 160035, 2016.

[16] Iris Eekhout, R Michiel de Boer, Jos WR Twisk, Henrica CW de Vet, and Martijn W Heymans, "Missing data: a systematic review of how they are reported and handled," *Epidemiology*, vol. 23, no. 5, pp. 729–732, 2012.

[17] Divyanshu Talwar, Aanchal Mongia, Debarka Sengupta, and Angshul Majumdar, "Autoimpute: Autoencoder based imputation of single-cell rna-seq data," *Scientific reports*, vol. 8, no. 1, pp. 16329, 2018.

[18] Adriana Fonseca Costa, Miriam Seoane Santos, Jastin Pompeu Soares, and Pedro Henriques Abreu, "Missing data imputation via denoising autoencoders: the untold story," in *International Symposium on Intelligent Data Analysis*. Springer, 2018, pp. 87–98.