ISBN: 978-981-18-3959-7
2022 the 12th International Workshop on Computer Science and Engineering (WCSE 2022)
doi: 10.18178/wcse.2022.06.055

Feature Completion Using Correlation-Preserved Autoencoder

Li Sun, Tao Liu and Jiyun Li +

School of Computer Science and Technology, Donghua University Shanghai, China

Abstract. Missing value existing in various datasets always causes tremendous obstacles to data mining in the real world. In recent years, autoencoder has been a popular deep learning model in data imputation due to the simple structure and efficient training period. In this paper, we developed a model that combined advantages of multiple imputations using denoising autoencoder (MIDA) and tracking-remove autoencoder (TRAE), integrating the idea of tracking-remove into MIDA. We introduce KNN to pre-impute the dataset after the input was denoising, and then we put the pre-imputed input data into both MIDA and MIDA with tracking-remove and ensemble the outputs by linear combination corresponding to the "multiple imputation" thought. The model called correlation-preserved autoencoder (CPAE) is applied to the completion of brain tissue feature data in ADNIMERGE (ADNI database). Experiments show that CAPE has a better performance than MIDA, TARE, and other autoencoders.

Keywords: data imputation; Alzheimer's disease; autoencoder; KNN; data fusion.

1. Introduction

Missing data is very common in real world such as: traffic data [1], medical data [2] etc. There are many reasons for the missing of data. For example, information is temporarily unavailable; data is corrupted due to human factors; data storage equipment is lost because of equipment failure, storage media or transmission media failure. The high costs of storing information are a big concern. The missing data can be categorized into three types: missing at random (MAR), the probability of data missing is not related to the lost data itself, but only to part of the observed data, the miss of data is not completely random, the deletion of this type of data depends on other complete variables; missing completely at random (MCAR), the values missed completely at random. This type of miss does not depend on any incomplete variable or complete variable and has no effect on the unbiasedness of the sample; missing not at random (MNAR), the missing parts correlate with the values of the incomplete variable itself [3].

In the current study, the commonly used imputing methods are mainly divided into statistic based missing value filling methods and machine learning-based filling methods. In the statistical filed, multiple imputation method based on different models or rules to generate several imputed values and generate several imputed datasets. After statistical analysis, final imputed values were calculated by a rule. Multiple imputations have higher accuracy than single imputation, but there is an enormous increase in the calculation. In the machine learning filed, the cluster-based imputation method uses a clustering algorithm to divide the samples into different clusters, fill the missing values with reference to the cluster center and the complete samples. Fuzzy C-Means (FCM) calculates the membership of samples to individual clusters, thereby improving more flexible clustering results. In neural network-based filed, models take the existing complete dataset as a training dataset to adjust the network parameters, and fill the missing term using the trained models. Multilayer Perceptron (MLP) organizes several neuronal nodes into a layer of a neural network, and forms a complex nonlinear system with activation functions and connection weights between nodes. This approach is able to fully exploit the correlation between attributes, but requires constructing an individual model for each of the missing types, so the training process is time-consuming.

+ Corresponding author. Tel.: + +021-67792293;
E-mail address: jyli@dhu.edu.cn.

2. Related Work

In this paper, we take autoencoder as an imputation model, which has an input layer consistent with the output layer. All kinds of missing data would be imputed by only one model. Compared with MLP, autoencoder has a more succinct structure. Lovedeep Gondara and Ke Wang proposed a multiple imputation model based on overcomplete deep denoising autoencoders. Multiple imputation using denoising autoencoders (MIDA) [4], which is capable of handling different data types, missing patterns, missingness proportions and distributions. Xiaochen Lai et al. proposed a tracking-remove autoencoder (TRAE) to solve the identity mapping problem existing in the traditional autoencoder [5]. We introduce KNN to pre-impute the dataset after the input was denoising, which preserves the correlations between features. After denoising, part of the input and output are not equal, resulting the output lose track action from input. Therefore, we put the pre-imputing input data into both MIDA and MIDA with tracking-remove and ensemble the outputs by linear combination corresponding to the multiple imputation thought. CPAE not only retains the feature correlations, but also solves the identity mapping problem in the autoencoder.

3. Methodology

The correlation-preserved autoencoder combines both advantages of MIDA and TRAE, which could shorten the training time and achieve better accuracy under various data missing patterns. The CPAE framework is demonstrated in Figure 1. In this section, we first give a detailed introduction to MIDA and TRAE. Then, we use KNN to preserve the feature correlations from denoised input and CPAE to solve the identity mapping issue.

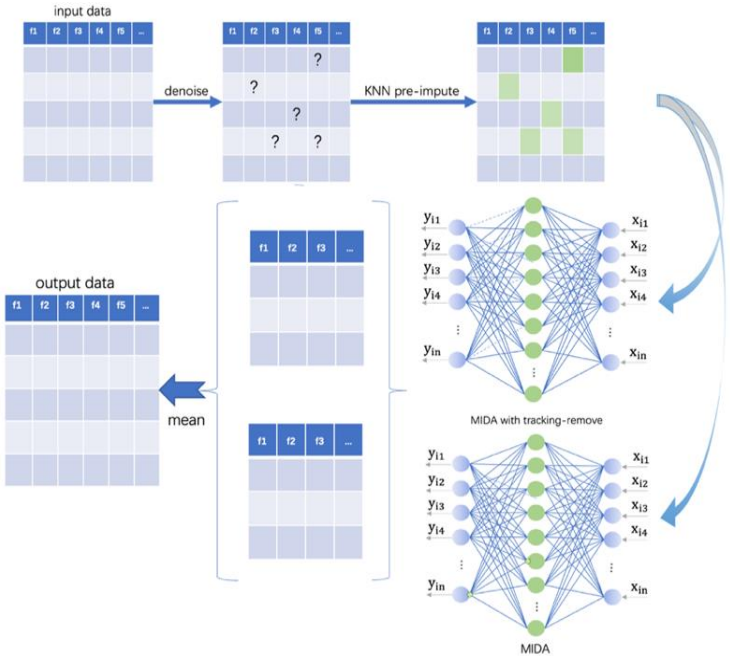

Fig.1: The framework of CPAE

3.1. MIDA

Denoising autoencoder (DAE) is an improved autoencoder designed to add noise to the input data to corrupt the identity map. In the DAE, part of input data is replaced by random values or zeros; the feature correlations hidden in the original data may be broken, resulting in a bad imputation. MIDA uses an overly complete DAE as a compensation model by mapping the input to a high-dimensional subspace, regaining the corrupted feature correlations during model training. MIDA uses multiple sets of random weights to initialize the model, obtaining multiple filling results. Finally, the mean value of the multiple imputation results is token as the imputation result.

The architecture of MIDA is shown in Figure 2, as we can see, original data inputs to autoencoder after denoising, H_1、H_2 ... H_6 represent hidden layers in encoder and decoder, more units in the successive layer during encoding phase compared to input layer for purpose of creating representations capable of adding lateral connections, aiding data recovery. Encoder and decoder are constructed using fully connected artificial neural networks. Input data will be input k times to get k results. The final imputation is calculated by the mean of k results.

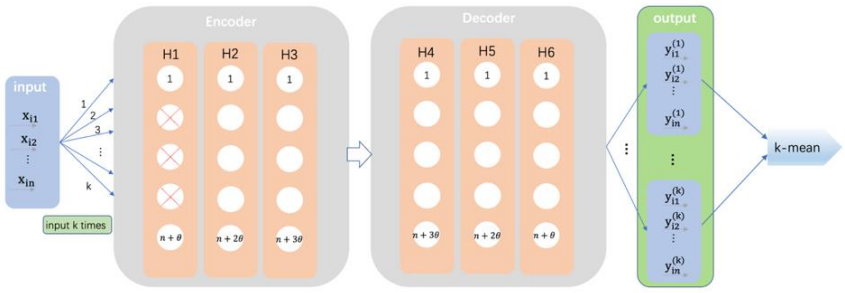

Fig.2: The architecture of MIDA

3.2. TRAE

In dealing with the medical dataset, the autoencoder can handle complex missing types and is efficient in training and imputing periods. During the training, the error between input and output constantly decreased, the output is easy to track the corresponding input, so the autoencoder achieves the training goal through a meaningless identity map, but the true correlations between attributes have not been learned by the model. The unique distinction is that TRAE modifies the calculation rules of the hidden layer on the basis of a normal autoencoder so as to weaken the identity mapping from the input layer to the output layer. Taking the hidden layer neuron of the kth in Figure 3, the difference between a normal autoencoder and TRAE is shown.

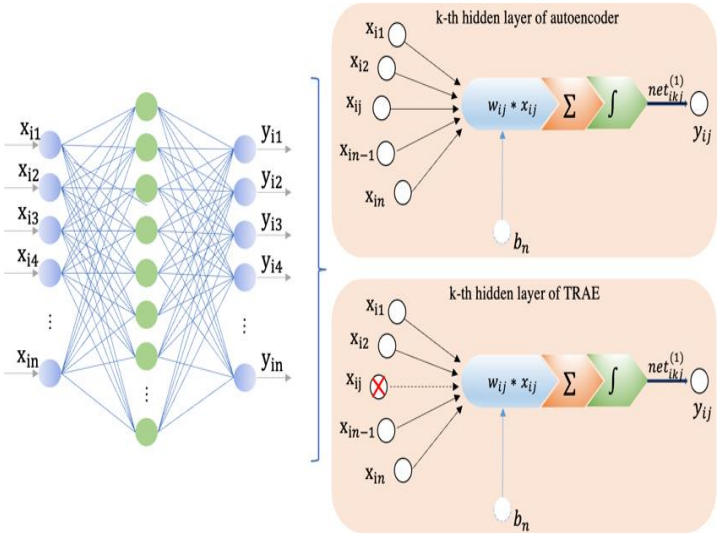

Fig.3: The architecture of TRAE and calculation rule in hidden layer neuron

In a normal autoencoder, the calculation rule in kth neuron of hidden layer is that: sum all inputs with thresholds and pass the result to the activation function. The expression is shown below.

$$net_{ikj}^{(1)} = \varphi\left[\sum_{l=1}^{n} w_{lk}^{(1)} * x_{il} + b_k^{(1)}\right], \quad \mathrm{j} = 1, 2, 3, \ldots, \mathrm{n} \tag{1}$$

While the TRAE calculation rule is that:

$$net_{ikj}^{(1)} = \varphi\left[\sum_{l=1,l\neq j}^{n} w_{lk}^{(1)} * x_{il} + b_k^{(1)}\right], \qquad \mathrm{j} = 1, 2, 3, \ldots, \mathrm{n} \tag{2}$$

Where $w_{lk}^{(1)}$ is the weight between lth input layer and kth output layer neuron. $b_k^{(1)}$ represents the threshold of the kth hidden layer neuron. $net_{ikj}^{(1)}$ is the output value passed in by the kth hidden layer neuron to the jth output layer neuron for the sample x_i. $\varphi()$ is the tanh function.

In a normal autoencoder, the $net_{ikj}^{(1)}$ that a particular hidden layer neuron passes into all output layers is exactly equal, while TRAE is different. In the improved hidden layer neuron, the input structure adjusts dynamically depending on the output object.

Through comparison, it can be found that x_{ij} participates in the solution of the y_{ij} as an independent variable, and the role of TRAE is to eliminate the influence of the x_{ij} on the result, but still let the model effectively reconstruct the $y_{ij} = x_{ij}$. In a normal autoencoder model, because of the existence of x_{ij}, y_{ij} relies heavily on the x_{ij}, resulting in ignoring the correlations between attributes.

3.3. CPAE

In the MIDA model, data denoising is ahead of input with a specific missing rate, part of the selected data is replaced with random values or zero values, as mentioned by the author in his paper, after the denoising process, the original data feature of the implied correlation will be erased, and finally, the data filling effect on the test dataset is not ideal. So the author proposes to increase θ neurons of each hidden layer by mapping the data attributes to high-dimensional space to obtain the interconnectivity between them. This improvement method has certain defects on many occasions. For example, if the missing rate is not large, the correlation between the properties of the denoising processing is not large, blindly increasing the number of layers of the hidden layer and the number of neurons will lead to long training time and overfitting; in the data set with a large amount of data, features with a high proportion of missing data or high correlation are replaced by a large number of random values during the noise reduction process, and it is pretty difficult to increase the depth of the hidden layer and the number of neurons to capture the correlation information.

In view of the above situation, we propose to fill the denoised data with null, and then the KNN is used to pre-fill the denoised data. The KNN looks for k complete samples that are closest or most relevant to each incomplete sample. Minkowski distance [6] is commonly used for distance measure. The distance $d_{\min}(x_i, x_j)$ is

$$d_{\min}(x_i, x_j) = \left[\frac{s\sum_{l=1}^{s} \mathrm{I}_{il} * I_{kl} * |x_{il} - x_{kl}|^p}{\sum_{l=1}^{s} I_{il} * I_{kl}}\right]^{\frac{1}{p}},$$
$$k = 1, 2, \ldots, n;\ k \neq i, \qquad \sum_{l=1}^{s} I_{il} * I_{kl} \neq 0 \tag{3}$$

After applying KNN to pre-impute denoised data, the correlations among features can be well preserved. In the subsequent step, the depth of the hidden layer and the number of hidden layer neurons can be reduced to achieve faster training when building the network model.

The autoencoder is inherently easy to learn the identity mapping of input and output. So adding tracking-remove thought to CPAE can effectively prevent this issue. In the above improvement method, the input data after pre-impute is not equal to the target data. Part of the data loses the meaning of identity mapping. Based on the idea of multiple imputations, CPAE combines MIDA and MIDA with tracking-remove to impute the missing values separately, and the final result is the mean value of these two outcomes, as Figure 1 demonstrates.

4. Experiment

4.1. Dataset

This dataset comes from the ADNIMERGE collection of CSV data in the ADNI database, which contains the basic information of the subject's sex, age, years of education, etc. The information of brain tissue

characteristics includes two diagnostic records: initial diagnosis, follow-up, feature information extracted from medical images, including voxel information such as hippocampus, brainstem, whole brain, temporal lobe, and so on. The dataset used in this paper is data of brain tissue features containing two diagnoses extracted from ADNIMERGE, because AD is a neurodegenerative disease, from mild cognitive impairment (MCI) to Alzheimer's disease is a gradual process. The addition of age and the timing of two diagnoses helps to strengthen the link between attributes. The data deletion type of the processed dataset is MAR, and the missing values can be filled by the relationship between the attributes.

4.2. Experiment Settings

The unit of measurement in ADNIMERGE brain feature is voxels, in each attribute, different features may have a large difference such as hippocampus and wholebrain, so the individual tissues in the voxel size range of 1000 to 180000, in order to eliminate the dimension, accelerate the optimization process, the experiment using MinMaxScaler to map the input data to [0, 1]. The imputation experiment is divided into two processes, first dividing the complete samples in ADNIMERGE, dividing the training set and the validation set into a 7:3 ratio for the complete sample, and after the model training is completed, the missing samples are entered into the trained model to complete the imputing task of the missing values.

The combination of the above parameters is used to train the model. After the model training is completed, the data on the test set is randomly deleted, and the data with missing values are entered into the model to obtain the predicted value of each missing term. The evaluation metrics are Root Mean Square Error (RMSE) and Mean Absolute Percentage Error (MAPE) for the effectiveness of the model.

MAPE is used to calculate the error between the imputing value and true value [7], as follows:

$$MAPE = \frac{1}{|\hat{X}_M|} \sum_{\hat{x}_{ij} \in \hat{X}_M} \left| \frac{r_{ij} - \hat{x}_{ij}}{r_{ij}} \right| \tag{4}$$

where $\hat{X}_M$ is the set of imputing values and r_{ij} represents the true value corresponding to the imputed value $\hat{x}_{ij}$.

RMSE is used to measure the sample standard deviation between the original data and the imputed data, with higher weights for larger errors [8]:

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2}, \in [0, +\infty) \tag{5}$$

where y_i is the true value, $\hat{y}_i$ is the imputed value and n is the number of missing items.

4.3. Experiment Result and Analysis

In order to prove that the CPAE model has better performance on AD feature completion, the experiment is divided into two dimensions for comparison, first of all, with MIDA, TARE, using KNN pre-imputed MIDA (KNN-MIDA-AE) and MIDA with tracking-remove (KNN-MIDA-TARE), and then combining the currently popular improved autoencoder such as denoising autoencoder (DAE), Multiple imputation Autoencoder (MAE) and Variational Autoencoder (VAE). The results are illustrated in Table 1 and 2, respectively. By analyzing the missing sample data to be filled, 40% of the data of the test set is processed to mimic the missing pattern.

Table 1: MIDA, TRAE and CPAE performance

model	**RMSE**	**MAPE**
MIDA	180934	52.47
TARE	135581	42.52
KNN-MIDA-AE	130215	42.03
KNN-MIDA-TARE	135070	47.60
CPAE	**95470**	**40.37**

Table 2: CPAE and other autoencoders performance

model	RMSE	MAPE
DAE	632993	252.13
MAE	729902	284.35
VAE	587732	230.24
CPAE	**95470**	**40.37**

From the above two tables, it can be seen that the performance of CPAE on the completion of AD brain tissue features is better than that of other autoencoders.

5. Conclusion

In this paper, we propose CAPE that integrates MIDA and TARE, which has significantly improved the effect on data imputation after combining KNN for pre-impute and data fusion. Alzheimer's disease from normal to mild cognitive impairment and disease is a very long term, brain features change in the study of early screening of Alzheimer's disease has important significance, CPAE proposed in this paper applied to the brain feature completion, has a greater effect on the subsequent study of Alzheimer's disease. The future work of this paper focuses on the use of the label of patient diagnosis results in ADNIMERGE, and the classification label is used in the process of imputing to improve the accuracy of the filling values, so as to achieve the effect of mutual promotion and checks and balances between classification and regression.

6. Acknowledgements

This work has been supported by the National Key R&D Program of China under Grant 2019YFE0190500, the Fundamental Research Funds for the Central Universities of Ministry of Education of China (Grant No.2232021D-22), and the Initial Research Funds for Young Teachers of Donghua University.

7. References

[1] LOPEZ-MARTIN M, CARRO B, SANCHEZ-ESGUEVILLAS A, et al. Network traffic classifier with convolutional and recurrent neural networks for Internet of Things [J]. *IEEE Access*, 2017, 5: 18042-50.

[2] WANG Y, WANG L, RASTEGAR-MOJARAD M, et al. Clinical information extraction applications: a literature review [J]. *Journal of biomedical informatics*, 2018, 77: 34-49.

[3] LITTLE R J, RUBIN D B. Statistical analysis with missing data [M]. *John Wiley & Sons*, 2019.

[4] GONDARA L, WANG K. Mida: Multiple imputation using denoising autoencoders; *proceedings of the Pacific-Asia conference on knowledge discovery and data mining*, F, 2018 [C]. Springer.

[5] LAI X, WU X, ZHANG L, et al. Imputations of missing values using a tracking-removed autoencoder trained with incomplete data [J]. *Neurocomputing*, 2019, 366(Nov.13): 54-65.

[6] ZHANG S. Nearest neighbor selection for iteratively kNN imputation [J]. *Journal of Systems and Software*, 2012, 85(11): 2541-52.

[7] AKIL Y S, MIYAUCHI H. Elasticity coefficient of climatic conditions for electricity consumption analysis; *proceedings of the International Conference on Power System Technology*, F, 2010 [C].

[8] KAMBLE V B, DESHMUKH S N. comparision between accuracy and mse, rmse by using proposed method with imputation technique article history keywords [J]. 2018.