Missing Data Imputation for Machine Learning

Shaoqian Wang, Bo Li, Mao Yang(✉), and Zhongjiang Yan

School of Electronics and Information, Northwestern Polytechnical University,
Xi'an, China
wangshaoqian@mail.nwpu.edu.cn, {libo.npu,yangmao,zhjyan}@nwpu.edu.cn

Abstract. The imputation of missing values in datasets always plays an important role in the data preprocessing. In the process of data collection, because of the various reasons, the datasets often contain some missing values, and the excellent missing data imputation algorithms can increase the reliability of the dataset and reduce the impact of missing values on the whole dataset. In this paper, based on the Artificial Neural Network (ANN), we propose a missing data imputation method for the classification-type datasets. For each record which contains missing values, we make a list of the values that can be used to replace the missing data from the complete dataset. Our ANN model uses the complete records as the train dataset, and selects the most appropriate value in the list as the final result based on the label categories of the missing data. In our experiments, we compare our algorithm with the traditional single value imputation method and mean value imputation method with the Pima dataset. The result shows that our proposed algorithm can achieve better classification results when there are more missing values in the dataset.

Keywords: Data imputation · Machine learning · Artificial Neural Network

1 Introduction

With the development of data mining technology and the rise of machine learning technology, various datasets are becoming more and more important. Although everyone wants to get perfect datasets, in the process of data collection, any problem, such as human error and machine failure, will affect the datasets, resulting in abnormal or even missing values. In fact, many large datasets contain missing values. A large number of missing values will reduce the reliability of the whole dataset. For some datasets which are difficult to collect, the missing values problem will make enormous loss. At the same time, the missing data will also bring huge challenges to the data processing and analysis process and have an impact on the results of the experiment.

B. Li et al. (Eds.): IoTaaS 2018, LNICST 271, pp. 67–72, 2019.
https://doi.org/10.1007/978-3-030-14657-3_7

Because it is difficult to ensure that there is no error in the whole process of data collection, so the imputation algorithms of missing values are very necessary. Aiming at the problem of the missing values in the datasets, many researchers have proposed various data-filling algorithms. The simpler way to solve the problem is to delete records which contain missing values, single value imputation algorithm and mean value imputation algorithm, but all these algorithms have great limitations. The rest of the filling algorithms can be divided into statistical imputation methods and data mining based imputation methods. Commonly used statistical imputation methods include EM imputation, regression model imputation, multiple erasing difference imputation, et al. Classical data mining methods include neural network, decision tree, Bayesian network, KNN algorithm, etc. These algorithms have been widely applied to various kinds of missing values problems according to their advantages and characteristics, and have achieved remarkable results.

In this paper, we propose imputation algorithm based on neural network for classification-type datasets. We use the complete data in the dataset as training dataset to train our ANN model. Our ANN model can show probabilities of a record being classified as different categories. For each record which contains missing values, we use the different appropriate values from other complete records to replace missing values. In this way, we get a list of new records and we will select the best one from these records as the imputation result. Based on our ANN model, We can achieve the probability of the records which are classified as the correct categories. The record which achieve the highest probability will be select. Compared with other data imputation algorithms, this method takes advantage of the characteristics of the classified dataset label, thus further improving the imputation effect in the classified dataset. In our experiment, compared with the single value filling and mean filling method with Pima datasets, our proposed method can achieve better classification results when there are many missing values in the dataset.

In the rest of this article, the second part introduces the background knowledge and some related work of this algorithm. The third part introduces the concrete implementation process of our algorithm. The fourth part explains the experimental process and result analysis based on Pima dataset. The fifth part summarizes the full text.

2 Related Work

In some datasets that contain very few missing data, ignoring method, deleting method and zero value method are often used for data filling [1,2]. These methods have an effective effect when the missing values have little effect on the dataset, but these methods are no longer applicable when there are more missing data in the dataset. In a variety of complex missing data problems, various data imputation algorithms based on machine learning have achieved good results.

The KNN method is often used in the data imputation algorithm, and the Batista [3] proposed the KNNI algorithm. For a record Ri containing the missing value, the KNN algorithm is used to find the most similar k records to Ri in the whole dataset, and then mend the missing values in Ri based on the value in k records. The processing of KNNI algorithm is simple, but when the dataset is large, it will take a long time to calculate.

Rahman [4] proposed an imputation method named DMI based on decision tree. For the record Ri which contains missing values, DMI algorithm will select all the records belonging to the same leaf node with the record Ri in the dataset as a new dataset Di. Based on the dataset Di the EMI algorithm will be used to finish the imputation work. The quality of decision tree model has great influence on the effectiveness of DMI algorithm. If the classification effect of leaf nodes is not good enough, the DMI algorithm will be affected.

Neural network algorithm is also widely used in data mining based incomplete data imputation problem. The main idea is to minimize the error between the simulated value and the actual value of the network output, and use the error to adjust the weight. The neural network algorithm has a strong generalization as its advantage.

Silvaramírez et al. [5] designs a 3 layer perceptual network for data filling. The number of neurons in the network input and output layer are equal to the number of attributes of the dataset. By artificially deleting some data, the disturbance dataset is generated as input, the original complete dataset is trained as the output for the network, and then the data is filled with the obtained network. Compared with most machine learning based filling algorithms, the method can get higher filling effect. But this method will take a long time to train the model, and it don't take advantage of the labels in the datasets.

In fact, there are few cases of labels missing in classification-type datasets. The label is an important attribute to judge the similarity between different records. In this paper, our proposed algorithm builds a classification model based on ANN, and takes advantage of different labels in dataset to complete data imputation. Although it can only be used with classification-type datasets, experiments show that this method can significantly improve the classification accuracy of incomplete classification datasets.

3 Data Imputation Method

The algorithm in this paper constructs an artificial neural network with three hidden layers and a Softmax classifier. Softmax classifier can show the probability of the records which are classified as different categories. The process is mainly divided into the training of the model and the selection of the appropriate filling value by the trained model.

3.1 Training Step

From the dataset which contain missing values, all the complete records are selected to form a new dataset as the training dataset. In this paper, in order to test the imputation effect of datasets on different data missing percentage, we artificially deleting some data randomly. All the records which have been deleted form the missing-value dataset Dm, the remaining complete records constitute the complete dataset D, then the dataset D is used as the training dataset to the train our ANN model, thus a model for classification is obtained. The process is shown in Fig. 1.

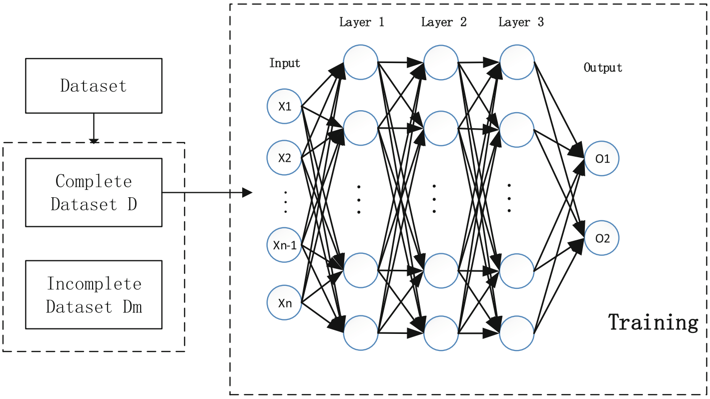

Fig. 1. ANN training with complete datasets

3.2 Data Imputation

After the network training is completed, our algorithm can solve the data imputation based on our ANN model. For a record Xi which have missing values, the first step in our method is find out the attributes which have missing values in this record, the values of these attributes from other complete records form the dataset Vm. The values in Vm will be used as the different imputation alternatives for the missing values of record Xi. The second step is to combine the various values in Vm with the complete part of the record. The missing values of the record Xi will be filled in respectively with all the values in Vm to form a dataset Dxi, which contain the final imputation result for the record Xi. In the third step, ANN model will calculate a list of outputs by the Softmax classifier using the records in Dxi. These outputs show the probabilities that the records are classified as different categories. In fact, the labels of the records in Dxi are the same with Xi's label Yi. The record Zi which achieve the highest probability to be classified as the category of Xi will be the best imputation result for the record Xi. By using the 3 steps above for all records in the incomplete dataset Dm, the final imputation dataset can be obtained. The whole process is shown in Fig. 2.

4 Evaluation

In the paper, we use the Pima Indians Diabetes to test the proposed method. This dataset contains the incidence of diabetes in hundreds of people within 5

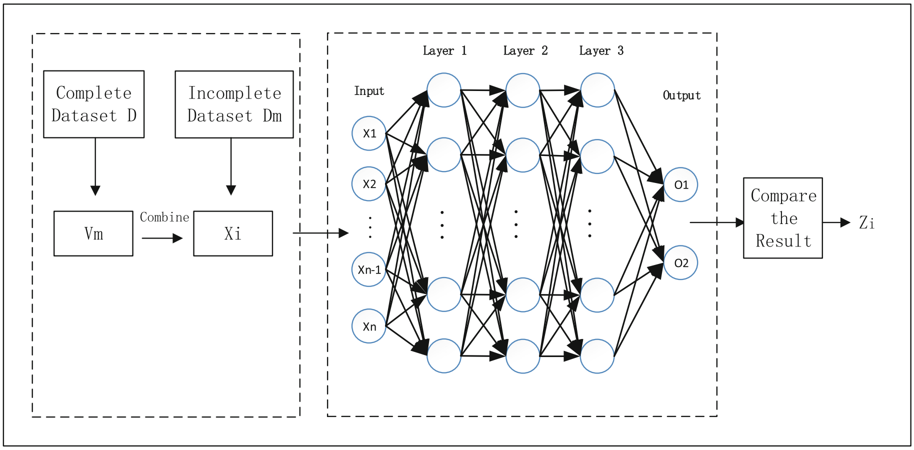

Fig. 2. Imputation process

years. It contains 8 types of physical condition variables for each person and 1 label which indicate whether a person is suffering from diabetes.

We first divide the whole dataset into training dataset and test dataset. We take 80% of the dataset as the training dataset, and the remaining 20% of the dataset as the test dataset. The Pima dataset is not a complete dataset, there have been some missing values in some attributes. We build an ANN model with 3 hidden layers, the number of neurons in each layer is 12, 24 and 16. Base on this ANN model, we achieved a result of 82.285% classification accuracy with our training dataset and test dataset. In order to test the classification results of datasets after data imputation with different missing rate, we delete the values in the datasets artificially, and get 3 new datasets which have 30%, 50%, 70% missing rate respectively. Each record in the new datasets contains a maximum of one missing value. Then, according to the imputation process described in the last section, we achieve 3 complete datasets after the imputation work. We achieve the classification accuracy of these 3 new complete datasets with our ANN model, and compare the results with the mean value imputation method and zero value imputation method. The mean imputation method replaces the missing values with the mean of the whole values of the missing values' attribute and the zero value method replaces all the missing locations with zero values.

Table 1. Classification accuracy with different imputation methods

Missing rate	Zero value method (%)	Mean value method (%)	Our method (%)
30%	80.12	80.56	84.96
50%	78.57	79.67	84.28
70%	77.14	78.42	84.20

From Table 1, we can find that with the increase of missing proportion, the classification accuracy of the 3 imputation methods has decreased. At the same time, through the lateral contrast, we can find that the classification accuracy of the zero value imputation is slightly higher than the mean imputation for the dataset. Meanwhile, the classification accuracy of our proposed method is greatly improved compared with the other two kinds of methods. For all the kinds of missing proportion, the accuracy of classification increased by more than 4%. Moreover, with the increase of data missing proportion, the accuracy of classification achieved by our method has not decreased a lot.

5 Conclusion

This paper proposes an artificial neural network-based incomplete data imputation method for categorical datasets. A classification neural network model is obtained by training with the complete records, and the missing values are fixed with possible values. The optimal imputation value is selected as the final imputation result by our ANN model. We test our method on the Pima dataset and compare the classification accuracy with zero value imputation method and mean imputation. The results show that our method has a greater improvement in the classification accuracy compared with zero imputation method and mean imputation method at different data missing degrees.

Acknowledgement. This work was supported in part by the National Natural Science Foundations of CHINA (Grant No. 61771390, No. 61771392, No. 61501373, and No. 61271279), the National Science and Technology Major Project (Grant No. 2016ZX03001018-004, and No. 2015ZX03002006-004), and the Fundamental Research Funds for the Central Universities (Grant No. 3102017ZY018).

References

1. Cheng, Y., Miao, D., Feng, Q.: Positive approximation and converse approximation in interval-valued fuzzy rough sets. Inf. Sci. **181**, 2086–2110 (2011)
2. Meng, Z., Shi, Z.: Extended rough set-based attribute reduction in inconsistent incomplete decision systems. Inf. Sci. **204**(20), 44–69 (2012)
3. Batista, G.E.A.P.A., Monard, M.C.: An analysis of four missing data treatment methods for supervised learning. Appl. Artif. Intell. **17**(5–6), 519–533 (2003)
4. Rahman, G., Islam, Z.: A decision tree-based missing value imputation technique for data pre-processing. In: The Australasian Data Mining Conference, pp. 41–50 (2010)
5. Silvaramírez, E.L., et al.: Missing value imputation on missing completely at random data using multilayer perceptrons. Neural Netw. Official J. Int. Neural Netw. Soc. **24**(1), 121–129 (2011)