Research on incomplete data mining and filling algorithm during depth learning process

Liping Wang
Pingxiang University, Pingxiang Jiangxi, 337055

Keywords: Depth learning; missing data filling; automatic coding

Abstract. In this paper, an incomplete data padding algorithm based on depth learning is proposed. The algorithm has a rich information dimension for large data. A depth-filling network is constructed to extract the depth features of large data, and then the missing values are restored. Experimental results showed that the algorithm proposed in this paper can effectively improve the accuracy of data filling. To solve this problem, this paper proposes an incomplete data filling algorithm based on depth learning. The algorithm is based on the automatic coding machine to establish the automatic filling machine. On this basis, a deep-filling network model is constructed to analyze the depth characteristics of incomplete data and calculate the network parameters according to the layer-by-layer training idea and back propagation algorithm. Finally, the incomplete data is restored by deep filling network, and the missing value is filled. In the next step, we explore how to improve the data filling accuracy in multi-miss mode.

Introduction

With the rise of Internet of Things, social networking and e-commerce technology, data is growing at an unprecedented rate [1, 2]. Large data research and application time has come. In the process of large data collection and transmission, each link may fail, resulting in a lot of large data sets there are a lot of missing [3]. Large data incompleteness poses great challenges for large data analysis and processing. Therefore, the filling of incomplete data is of great significance to the analysis and processing of large data [4].

In recent years, domestic and foreign researchers have proposed a lot of incomplete data filling method. Including data packing algorithm based on maximum expectation, filling algorithm based on fuzzy clustering and filling algorithm based on nearest neighbor object [5]. These algorithms have achieved remarkable results in filling small-scale data sets. However, in filling incomplete data, the accuracy of sharp decline [6]. This is due to the large data there is a rich information dimension, and the traditional data filling algorithm does not reflect the depth characteristics of large data. Aiming at this problem, this paper proposes an incomplete large data padding algorithm based on depth learning [7, 8].

In this paper, an incomplete data padding algorithm based on depth learning is proposed. The algorithm has a rich information dimension for large data, and a depth-filling network is constructed to extract the depth features of large data, and then the missing values are restored. Experimental results show that the algorithm proposed in this paper can effectively improve the accuracy of data filling. To solve this problem, this paper proposes an incomplete data filling algorithm based on depth learning. The algorithm is based on the automatic coding machine to establish the automatic filling machine.

1. Fill the automatic encoder

Let the entire set of data objects be O, and O = $\{O_1, O_2, \ldots, O_k\}$, each object in set O consists of m attributes $\{a_1, a_2, \ldots, a_m\}$, which is A={ a1, a2, …, am } is a collection of properties. The algorithm

first divides the entire data set into two sets C and U, which $C=\{c_1,c_2,\cdots,c_l\}, I=\{i_1,i_2,\cdots,i_p\}$. C is the complete dataset, that is, all the data objects in C are complete; I is a non-complete data set, that is, all data objects in I have missing attribute values.

In this paper, a deep filling network is constructed to fill the auto coder as the base module, and a part of the data objects is randomly selected from the complete data subset as an example to train the network parameters of the automatic encoder. In the process of constructing the automatic filling machine, the selected data object is used to simulate the missing data object. The partial attribute value of each instance data object is randomly set to 0, the incomplete object is simulated as the input of the filling automatic encoder, Constructs the data and the instance prototype to train the network parameter. Examples of automatic encoding machine were shown in Figure 1.

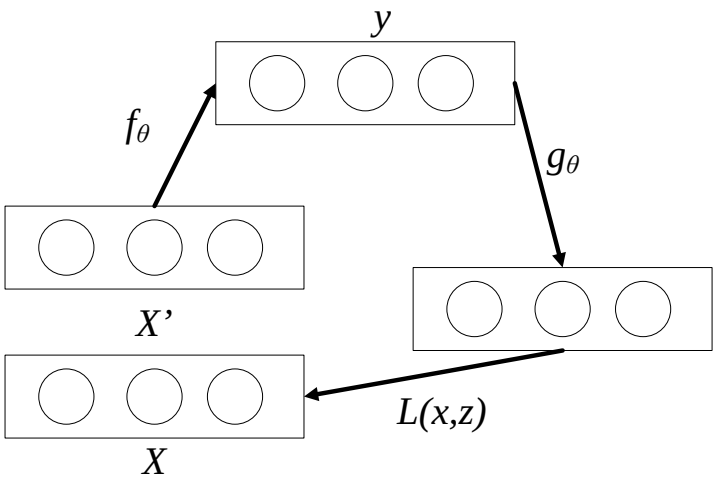

Figure 1. Fill the automatic encoder structure

Suppose n complete data objects are selected to constitute an instance set:

$$I=\{xi|\ i=1,2,3,\ldots,n\} \tag{1}$$

The model parameters are trained by minimizing the average reconstruction error such that the reconstructed data object tries to approximate the instance object as much as possible.

$$\begin{aligned}\theta^*,\theta'^* &= \arg\min_{\theta,\theta'}\frac{1}{n}\sum_{i=1}^{n}L(x_i,z_i)\\ &= \arg\min_{\theta,\theta'}\frac{1}{n}\sum_{i=1}^{n}L\left(x_i,g_{\theta'}\left(\left(f_\theta\left(x_i^{'}\right)\right)\right)\right)\end{aligned} \tag{2}$$

Consistent with the automatic encoder: $\theta=\{W,b\}$ *and* $\theta'=\{W',b'\}$ is network parameter.

According to the stochastic gradient descent algorithm, every time an instance is selected from data set I for training, the filling auto-coder first randomly selects some attributes of the instance and sets its attribute value to zero to get a pair of data x and x ' The weights of the automatic encoder are updated once by the following formula. This updates the network parameters until the entire network is stabilized.

$$\begin{aligned}W' &= W-\eta\left(\frac{\partial L(X,Z)}{\partial W}+\lambda W\right)\\ b' &= b-\eta\left(\frac{\partial L(X,Z)}{\partial b}+\lambda b\right)\end{aligned} \tag{3}$$

2. Deep fill network with data filling

Deep filling network and data filling in this paper, a three-tier network model is built up based on the auto-encoder-based module. Each layer of network output will be the input of the upper layer network, and the top layer will be output as the extracted feature. Training process is divided into

pre-training and fine-tuning two stages. First, the network initialization parameters are obtained from bottom to top, and finally the global parameters are fine-tuned by back propagation algorithm. In order to get the object of network layer-by-layer training and supervision, firstly, the example data is used as the input to construct the overlaying automatic coding machine. Overlay automatic encoder structure was shown in Figure 2.

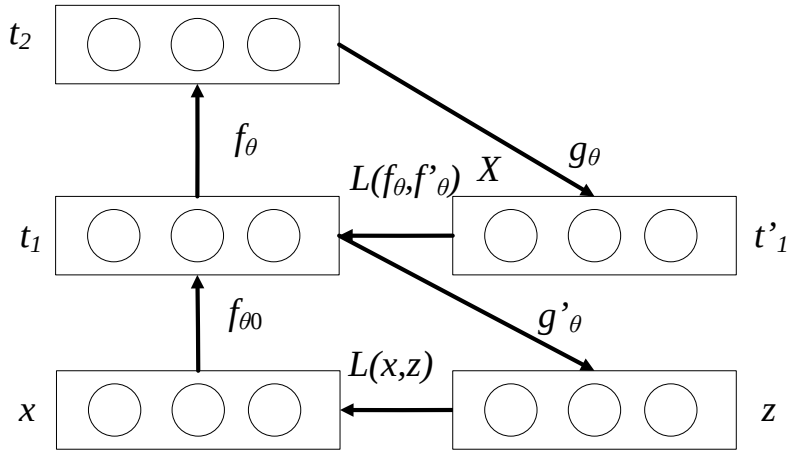

Figure 2. Two-level stacking automatic encoder structure

In this paper, the raw data x is taken as the network input, and the first layer feature t1 = fθ(x) is obtained at the lowest level, and the feature t1 is taken as the input of the upper layer network to obtain the second layer characteristic t2 = f θ(t1), the training process is local, that is, the second layer network to update the layer of the network weight, the lower the network has no effect. In this way, you can initialize the stack network parameters, and finally through the back propagation algorithm to fine tune the network global parameters. Thus, two-layer features t1 and t2 corresponding to the original data instance can be obtained.

3. Experiment and analysis

In order to verify the validity of the algorithm proposed in this paper, the algorithm proposed in this paper is compared with two filling algorithms, FIMUS and DMI. The data set used in this paper is collected from the digital home and wireless sensor network laboratory, the total data set to 10G each data object contains 650 numerical attributes. We first delete some data from the data set, and simulate incomplete data sets. After the filling is complete, we compare the filled values with the real values to get the filling accuracy of the algorithm.

In this paper, two missing values, single mode deletions and multiple mode deletions were artificially created. In single-schema misses, each data object is allowed to contain only one missing value, and multiple data misses allow each data object to have multiple missing values. In this paper, we select 1%, 3%, 5% and 10% data objects from the dataset and delete some of these data objects to simulate the missing data. This paper uses two criteria to measure the filling accuracy of the algorithm. The first standard is called the d_2 standard, which is used to measure the degree of match between the fill value and the true value. The formula is as follows:

$$d_z = 1 - \left[\frac{\sum_{i=1}^{n} (e_i - r_i)^2}{\sum_{i=1}^{n} \left(|e_i - E| - |r_i - R| \right)^2} \right] \tag{4}$$

The second criterion is RMSE, which measures the average error between the fill value and the true value, calculated as:

$$RMSE = \left(\frac{1}{n} \sum_{i=1}^{n} |r_i - e_i|^2 \right)^{1/2} \tag{5}$$

As can be seen from Table 1, for any combination of missing, the proposed algorithm d2 are significantly higher than the other two algorithms. In addition, with the increase of data missing rate, the algorithm FIMUS and DMI are decreased. The filling accuracy of these two algorithms decreases with the increase of data missing rate. The filling accuracy of the algorithm proposed in this paper has been maintained at a very high level. Thus, the filling accuracy of the proposed algorithm is significantly higher than that of FIMUS and DMI.

Table 1. Filling accuracy index:d2

Combination		Algorithm		
Loss Rate/%	Loss Mode	DLDBI	FIMUS	DMI
1	single	0.876	0.717	0.791
	multiple	0.838	0.668	0.709
3	single	0.879	0.750	0.629
	multiple	0.818	0.644	0.627
5	single	0.848	0.701	0.768
	multiple	0.845	0.651	0.651
10	single	0.871	0.778	0.762
	multiple	0.864	0.739	0.650

For any missing combination, we choose different objects as training data, run the algorithm 20 times, and get the mean value of d2 and RMSE, as shown in Fig.3. It can be seen from Figure 3, the proposed algorithm filling accuracy is relatively stable. Specifically, when the data loss rate is 1% to 10%, d2 values can be stably maintained above 0.8, RMSE values stable between 0.15 and 0.2.

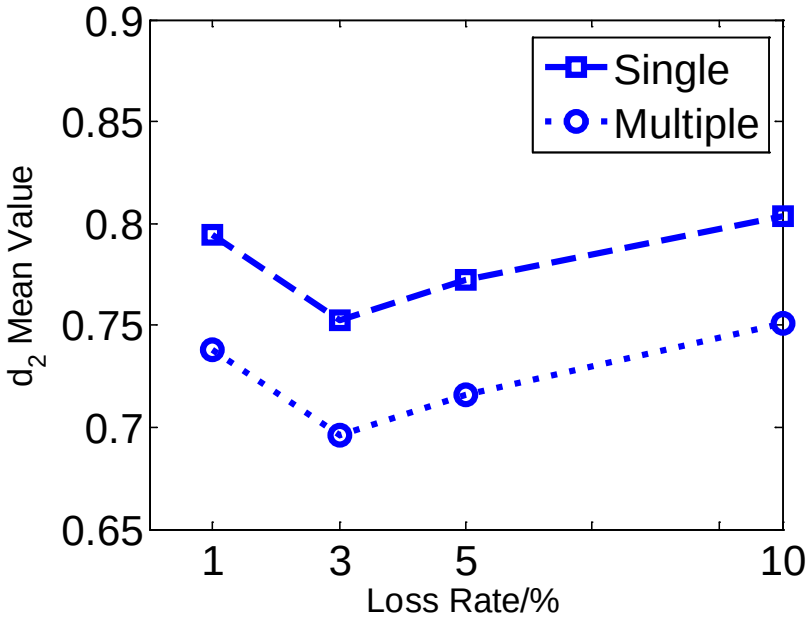

Figure 3. Average of d2 parameter

Summary

In this paper, an incomplete data padding algorithm based on depth learning is proposed. The algorithm has a rich information dimension for large data, and a depth-filling network is constructed to extract the depth features of large data, and then the missing values are restored. Experimental results show that the algorithm proposed in this paper can effectively improve the accuracy of data filling. To solve this problem, this paper proposes an incomplete data filling algorithm based on depth learning. In the next step, we explore how to improve the data filling accuracy in multi-miss mode.

Acknowledgement

This work was supported by Science and Technology Research Project in Department of Education, Jiangxi Province, 2015 (GJJ151274), Jiangxi Provincial Intellectual Property Soft Science Research Project (ZR201610), Humanities and Social Sciences Research Project of Jiangxi Province, 2015 (TQ1516),Art Science Planning Project of Jiangxi Province (YG2014255, YG2015189) and Pingxiang Science and Technology Support Program, 2015: Research and Development of Animation Rendering Service Platform in the Central Region Based on Cloud Computing.

.

References

[1] Huang F, Liu C, Huang Y, et al. Dynamic Cost-Sensitive Extreme Learning Machine for Classification of Incomplete Data Based on the Deep Imputation Network[J]. International Journal of Database Theory and Application, 2016, 9(6): 285-298.
[2] Q. H. Spencer, A. L. Swindlehurst, M. Haardt. “Zero-Forcing Methods for Downlink Spatial Multiplexing in Multiuser MIMO Channel”, IEEE Transactions on Signal Processing, vol.52, no. 2, pp.461-471, May 2004.
[3] Fournier H, Kop R, Durand G. Challenges to research in MOOCs [J]. Journal of Online Learning and Teaching, 2014, 10(1): 1.
[4] Xiong X, Adan A, Akinci B, et al. Automatic creation of semantically rich 3D building models from laser scanner data[J]. Automation in Construction, 2013, 31: 325-337.
[5] J.van de Beek,O. Edfors, M. Sandell, S. Wilson, P. Borjesson, “On Channel Estimation in OFDM System”, in Proceedings of the IEEE Vehicular Technology Conference, pp. 815-819, 1995.
[6] Pattanodom M, Iam-On N, Boongoen T. Clustering data with the presence of missing values by ensemble approach[C]//2016 Second Asian Conference on Defence Technology (ACDT). IEEE, 2016: 151-156.
[7] Liu H, Motoda H. Feature selection for knowledge discovery and data mining[M]. Springer Science & Business Media, 2012.
[8] Piech C, Sahami M, Koller D, et al. Modeling how students learn to program[C] // Proceedings of the 43rd ACM technical symposium on Computer Science Education. ACM, 2012: 153-160.