The Eighteenth International Symposium on Artificial Life and Robotics 2013 (AROB 18th '13),
Daejeon Convention Center, Daejeon, Korea, January 30-February 1, 2013

An estimating method for missing values by using multiple SOMs

Yuui Kikuchi[1], Nobuhiro Okada[1], Yasutaka Tsuji[1], and Kazuo Kiguchi[1]

[1]Kyushu University, Japan
(Tel: 81-92-802-3167)

[1] 2TE11313R@s.kyushu-u.ac.jp

Abstract: Recently, development of information technology grows the importance of data analysis. In actual data, however, instances will sometimes miss some of their values. Then, how to deal such missing values has become one of the important subjects. Estimating and completing the missing values is required when analyzing the instances or attributes including the values. Using a self-organizing map (SOM) is one of such estimating method. This method is available for nonlinear data. In the data which lacks instances which are not including missing values, however, it was difficult to estimate such values by using conventional SOM method. To solve the problem, we propose a new method that uses multiple SOMs for estimating missing values. To evaluate our method, we performed simulation using proposed and other conventional methods. By the simulation results, we showed the advantages of our method.

Keywords: Self-Organizing Map, Data Processing, Missing Value

1 INTRODUCTION

In recent years, collection and accumulation of various data becomes very easy because of development of information technology. Due to this change, the importance of data mining is increasing. Data mining is the work finding the knowledge hid in data. In actual data, however, instances will sometimes miss some of their values for various reasons. Then, how to deal such missing values has become one of the important subjects in data analysis. In general, there are two methods to deal with them. One is the removing instances or attributes including the missing values, and another method is estimating and completing them. The latter method is required when analyzing the instances or attributes including the missing values. Such a method is also applicable to data prediction problems. In this method, at first, the features or the laws of whole data are led from a part of data except for the missing values. Then, the values are estimated basing on the features or the laws.

Prior research on missing values is manifold. Edgar [1] investigated removing method and some estimating method. He investigated their influence in data analysis also. Abe et al [2] compared some statistical-analysis techniques to estimate missing values. Using a self-organizing map (SOM) is one of estimating method. For example, Iwasaki et al [3] studied application SOM to estimating missing values in earthquake data.

The SOM is an artificial intelligence which is proposed by Kohonen [4], and it is applied to data analysis and various field [5]. This method is also available for nonlinear data. The SOM method is very useful, because actual data is frequently nonlinear. The same can be said about missing values estimation. In conventional SOM method, however, only instances which are not including missing values can be used for learning the features or the laws of whole data. Thus, in the data which lacks such instances, it is difficult to estimate missing values by using SOM.

In this paper, to solve the problem, we propose a new method that uses multiple SOMs for estimating missing values. This method can achieve effective estimation even for data which lacks instances not including missing values. Moreover, the system is still available for nonlinear data because of using SOMs. To evaluate our method, we performed simulation of estimating the missing values with incomplete data. From the simulation result, we proved that our method is more effective than some conventional method including a conventional SOM.

2 THE ESTIMATING METHOD BY SOM

In this section, we describe the method to estimate missing values by using conventional SOM. We explain SOM in the sections 2.1-2.3, and the estimating method in the section 2.4.

2.1 A summary of SOM

SOM is the algorithm which makes a map reflecting the features of high-dimensional input vectors. The map consists of neurons latticed in low-dimensional space. Thus, the position of the neuron on the map is expressed with the coordinates in low-dimensions. Each neuron has a reference vector. The reference vectors have same dimensions as input vectors. In algorithm of SOM, the reference vectors

are updated, whenever an input vector is presented. By repeating updating, a map gradually learns the feature of input vectors. After sufficient learning, the map reflects the features.

2.2 The algorithm of SOM

The algorithm of SOM is described below. For convenience, a number is assigned to each neuron and *i*-th neuron's reference vector is expressed as $\boldsymbol{m}_i$.

Each neuron is initialized by given an initial value to a reference vector. The value is determined at random or based on the result of principal component analysis (PCA).

When the vector $\boldsymbol{x}(t)$ is given as *t*-th input (*t*=1, 2...), a reference vector is updated in the following procedures.

$$||x(t)-m_c(t)|| = \min_i ||x(t)-m_i(t)|| \quad (1)$$

$$m_i(t+1) = m_i(t) + h_{ci}(t)\cdot\{x(t)-m_i(t)\} \quad (2)$$

At fast, a winning neuron which has the reference vector with the greatest similarity to $\boldsymbol{x}(t)$ is chosen from all neurons. The number of a winner is set to c. Then, $\boldsymbol{m}_i(t)$ is updated by equation (2). By updating, $\boldsymbol{m}_i$ is more similar to $\boldsymbol{x}(t)$. $h_{ci}(t)$ is the function of *t* and l_{ci} which is distance in the map from c-th neuron to i-th neuron. An example of $h_{ci}(t)$ using a Gaussian function is shown below.

$$h_{ci}(t) = \alpha(t)\cdot\exp[-l_{ci}^2/\{2\sigma^2(t)\}] \quad (3)$$

$\boldsymbol{m}_i$ of the neuron near a winner on the map changes a lot. As learning proceed, $\alpha(t)$ and $\sigma(t)$ is gradually decrease. In the early stages of learning, a reference vector changes a lot in the wide range on the map. The range and amount of change decrease as learning proceed.

2.3 The map after learning

In this section, we describe the map after learning. To each input vector, a neuron with reference vector which is very similar to it exists. When some two neurons are located close together on the map, their reference vectors are also well alike. Thus, with the position change on the map, the reference vector of the neuron also changes smoothly. By mapping the each input vector to the winning neuron, the relation of inputs in high-dimensional space can be expressed with the coordinates in low-dimensions on the map. The law of each component of input vector is expressed with all neuron's reference vector discretely whether the law is linear or not.

The example of a map after learning is shown in Fig. 1.The features or the laws of two-dimensional input vectors are reflected in the one-dimensional map. The axes in the figure correspond to each component of an input vector (or reference vector). A Fig. 1 shows that the map after learning has the above-mentioned character.

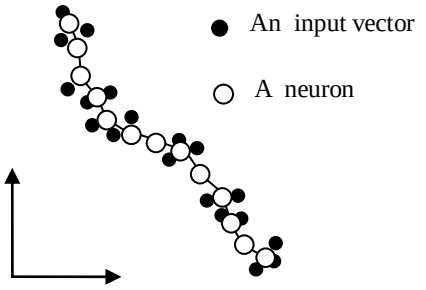

Fig. 1. The input vectors (2 dim) and the map (1 dim)

2.4 The estimating method by conventional SOM

In the method estimating missing values in data by SOM, at fast, instances which are not including missing values are used as input vectors for learning. The laws of each attribute of these instances are learned as the laws in all instances. After learning, the laws expressed with neurons of the map discretely.

Then, to each instance including missing values, the winner is determined. The winner is the neuron which has the reference vector with the greatest similarity to the instance. In the calculation of similarity, the attribute with missing value is excluded from consideration. It is the same also about the corresponding component of the reference vector. After determining the winner to each instance, missing value included in the instance is estimated. The estimated value is the corresponding component of a winner's reference vector.

In this method, however, only instances which are not including missing values can be used for learning. Thus, in the data which lacks such instances, it is difficult to estimate missing values by using SOM.

3 PROPOSED METHOD

In proposed method, a SOM disregarding some attributes are utilized. The SOM learn the relation between only attributes which is not disregarded. In the SOM, the instances including missing values can be used for learning, so long as the attribute with the missing value is disregarded. Thus, the SOM can use more instances than normal one when estimating missing values in data. On the other hand, the SOM can't learn the relation between the disregarded attribute and other attributes. Therefore, the SOM can't estimate missing values in the disregarded attribute.

We propose the method using multiple disregarding SOMs. The attribute disregarded by a SOM is not disregarded by other SOM. In multiple SOMs method, the relation of all the attributes is learned by the whole of multiple SOMs.

After learning, a missing value is estimated by each SOM. At this time, to one missing value, some estimated values are obtained from some maps. The average of these estimated values is used as a final estimated value.

The concrete way of using the proposed method is explained in section 4.4.

4 SIMULATION

4.1 A summary of simulation

To evaluate the proposed method, we performed simulation of estimating the missing values with incomplete data. In the simulation, the data for benchmark of data analyzing was utilized. The data had no missing values originally. And then we made the data incomplete by removing some values. We applied proposed method, conventional method that uses SOM and other methods to the incomplete data, and compared the results. We evaluated the results based on the average of each error of the estimated value to the original value. The error is a relative error to variance of the attribute with missing value.

In the algorithm of SOM in this simulation, we determined the initial value of reference vectors of neurons based on the result of principal component analysis (PCA). We used equation (3) as $h_{ci}(t)$ when updating a neuron.

4.2. The conventional techniques other than SOM

As the conventional technique for comparing with the proposed one, we adopted the method which used average value, and hot-deck imputation. In the method using average, the average value of an attribute with a missing value is adopted as an estimated value. In the latter method, at fast, the instance with no missing which is most similar to the instance with missing value is determined. In the calculation of similarity, the attribute with missing value is excluded from consideration. After determining the instance, the value which is corresponding missing values is adopted as estimated value.

4.3 The data for validation

In this simulation, we used the iris data which is released as a benchmark of analyzing [6]. The data has 150 instances and 4 attributes (attribute1-attribute4). The data had no missing values originally.

We made the data incomplete by removing some values. At fast, we determined the rate X(%) of the instances including missing values, and then, equally divided the instances into 4 groups at random. In each group, the value of the attribute 1, the attribute 2, the attribute 3, and the attribute 4 was made missed, respectively. These groups call Group1, Group2, Group3, and Group 4, respectively. The group of instances which are not including missing value calls Group0.

We created four kinds of incomplete data (Data1-Data4) by grouping the instances four times to each X (= 10, 20, …, 90).

When X = 90, to each incomplete data, we applied each estimating method, and calculated the average of the error, respectively. Furthermore, we compared proposed method and conventional SOM method when X = 10, 20, …, 90.

4.4 The estimation by proposed method

To estimate missing values by proposed method, we prepared four kinds of disregarding SOMs (SOM-A, SOM-B, SOM-C, and SOM-D) to each group of instances including missing value. SOM-A, SOM-B, SOM-C, and SOM-D disregards attribute1, attribute2, attribute3, and attribute4, respectively.

Then, each SOM learns the relation between attributes which are not disregarded. Each SOM can use more instances for learning than conventional one. For example, SOM-A can use not only instances of Group0, but also instances of Group1 which are including missing values in the attribute1 disregarded by SOM-A.

After learning, each SOM estimate a missing value in the attribute which are not disregarded. For example, the missing value of attribute3 is estimated by SOM-A, SOM-B, and SOM-D which don't disregard attribute3. The average of these estimated values is used as a final estimated value.

4.5 The simulation result

In the following Table 2, we show the simulation result when X = 90. The table 2 shows that the error of the proposed method (multiple SOMs) is the smallest.

In the following Fig. 2, we show the simulation result of estimation by the proposed method (multiple SOMs) and the conventional SOM method when X = 10, 20, …, 90. The Fig. 2 shows that the method using multiple SOMs is more effective than conventional SOM when X is at least 60(%).

Table 1. The error of estimated value when X = 90

	Data1 (%)	Data2 (%)	Data3 (%)	Data4 (%)	Ave (%)
Multiple SOMs	**32.8**	**36.7**	**35.5**	**31.6**	**34.1**
Conventional SOM	40.3	48.2	56.9	42.6	47.0
Using the average	87.5	88.1	81.7	83.1	85.1
hot-deck imputation	50.0	50.9	54.0	42.8	49.4

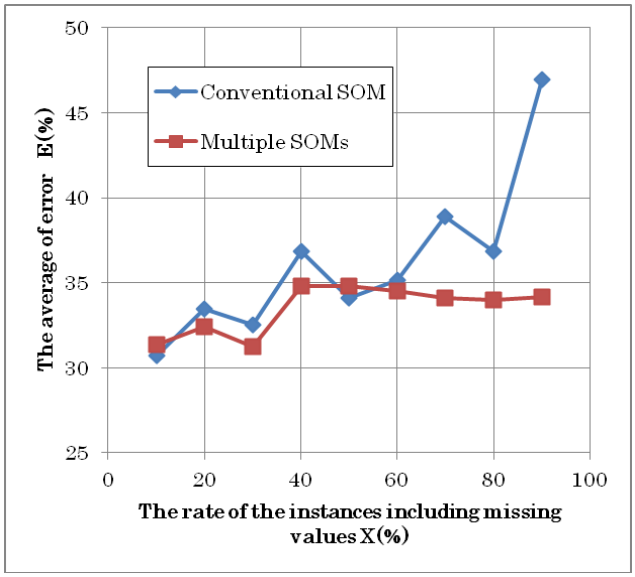

Fig. 2. The input vectors (2 dim) and the map (1 dim)

5 CONCLUSION

From the simulation result, we proved that our method is more effective than some conventional method including a conventional SOM especially in data with many instances including missing values.

In future, we have to perform more simulation in various data and conditions. For example, we have to check whether the proposed method has advantage even if the missing pattern becomes more complex. Moreover, we would like to consider any better method to determine a final estimated value based on some estimated values from each SOM.

REFERENCES

[1] Edgar A, and Caroline R (2004), The Treatment of Missing Values and its Effect on Classifier Accuracy. Classification, Clustering, and Data Mining Applications, Vol. 0, No.7, pp. 639-647

[2] Abe T, Inaba Y, Iwasaki M (2005), The Statistical-Analysis Techniques and Comparison of Software of Incomplete Data (in Japanes). Japanese Society of Computational Statistics, Vol. 18, No. 2, pp. 79-94

[3] Iwasaki T, Genei M (2001), Estimation of the earthquake information with missing by use of a self-organizing map (in Japanese). Technical paper summaries of Architectural Institute of Japan, B-2, p.185

[4] Kohonen T (2005), Self-Organizing Map (in Japanese), Springer-Verlag, Tokyo.

[5] George K (2010) Self-Organizing Maps, Publication In-Tech.

[6] University of California Irvine, Iris Data Set, http://archive.ics.uci.edu/ml/datasets/Iris