Estimation of Missing Values Using a Weighted K-Nearest Neighbors Algorithm

Wang Ling
Information Engineering School
University of Science and Technology Beijing
Beijing, China
linda_gh@sina.com

Fu Dong Mei
Information Engineering School
University of Science and Technology Beijing
Beijing, China
Fdm2003@163.com

***Abstract*—This paper developed a novel method to estimate the values of missing data by the use of a weighted -nearest neighbors algorithm. A weighting scheme that exploits the correlation between a "missing" dimension and available data values from other fields, which is quantified based on the support vector regression method. The proposed method has been applied to a practical case of modeling steel corrosion. Comparing with the traditional imputation algorithm, the model results demonstrate its better generalization capability.**

Keywords- missing values; SVR; weight; k-Nearest Neighbo; steel corrosion

I. INTRODUCTION

Missing values are a common problem in many data mining applications. There are many possible explanations for why a data value may be unavailable: the measurements were simply not made, human or machine error in processing a sample, and error in transmitting or storing data values into their respective records. When missing data are encountered, the simplest way of dealing with missing values is to discard the examples that contain the missing values. However, this method is practical only when the data contain relatively small number of examples with missing values and when analysis of the complete examples will not lead to serious bias during the inference. Another solution, more sophisticated one is to try to predict missing values with a data mining tool. In this case predicting missing values is a special data mining prediction problem.

The general methods have been mainly used for estimating multiple variables missing values in statistical analysis such as mean imputation [1,2,3], regression imputation[4], neural network imputation[5], category imputation[6]. The fourth approach for estimating missing values is more effective than other methods, of which K-Nearest-Neighbors(KNN)imputation method[7] is the most widely used method that the k data samples similar to the missing data samples were found to estimate the corresponding variable value of the missing variable value. However, the classical KNN imputation method ignores the statistical correlation that may be present between different attributes of a sample.

In this paper, we present a novel algorithm that is capable of simultaneously estimating several missing components using a weighted K-Nearest-Neighbors algorithm. The weights are quantified by the Support Vector Regression (SVR). SVR is a new promising regression technique based on the statistical learning theory. For a given learning task, with a given finite amount of training data, the best generalization performance will be achieved to model nonlinear system. So it is a good way used with KNN imputation to estimate the missing data.

II. KNN IMPUTATION

As an imputation approach, the K-Nearest Neighbors algorithm is very efficient and easily to be realized. In this method, the missing values of a sample are imputed considering k samples that are most similar to the sample of interest. The similarity of two samples is determined using a distance function. The algorithm is as follows:

1) Divide the data set D into two parts. Let Dm be the set containing the samples in which at least one of the variables is missing. The remaining samples will complete variable information form a set called Dc.

2) For each vector x in Dm:

(a) Divide the missing sample vector into observed and missing parts.

(b) Calculate the distance between the missing sample and all the sample vectors from the set Dc. Use only those variables in the sample vectors from the complete set Dc, which are observed in the vector x.

Let $x_j = [x_{j1,} \cdots, x_{jp}]$ has missing values, $x_i = [x_{i1,} \cdots, x_{ip}]$ denotes the complete data. This paper defines a distance metric between data points in inputs space, which is calculated as:

$$d_{ij} = \sqrt{(\sum_{l=1}^{p} (x_{il} - x_{jl})^2} \qquad (1)$$

Let d_{ij} gives the distance between x_i and x_j, x_{il} represents the value of variable l in sample x_i. The father distance is, the greater the difference between the samples is. The nearer distance is, the smaller the difference between the samples is. So, the similarity s_{ij} denotes the distance difference.

$$s_{ij} = 1 / d_{ij} \qquad (2)$$

(c) Calculate the weight of k samples vector.

978-0-7695-3682-8/09 $25.00 © 2009 IEEE
DOI 10.1109/ESIAT.2009.206

$$q_{ij} = \frac{s_{ij}}{\sum_{i=1}^{k} s_{ij}} \quad (3)$$

q_{ij} represents the weight of sample x_i corresponding to the missing data sample x_j.

(d) Use the k closest samples vectors (K-Nearest Neighbors) and perform a weighed estimate of the missing values for categorical attributes.

$$x_{jt} = \sum_{i=1}^{k} q_{ij} x_{it}, (t = 1, \cdots, p) \quad (4)$$

x_{jt} denotes the missing value of variable t in sample x_j.

(e) Repeat (d) for multiple variables fill in the missing value.

III. WEIGHTED-KNN IMPUTATION WITH MISSING VALUES

The KNN algorithm appears to be simple but powerful non-parametric imputation method. There is one problem, However, the distance calculation for how close a sample is to this sample is equally based on all of the present dimensions, which would make the estimation accuracy of the missing values decline. For example, every sample has 10 input variables, but only two input variables have a strong correlation to the output, three input variables have a weak correlation to the output, the other five input variables has nothing to do with the model output. In this case, if some attribute variable is highly correlated with a missing dimension, we should weigh the attribute variable higher in the distance between two samples, x_i and x_j. Therefore, in our distance metric, when imputing a missing dimension, we should weight each dimension by the respective correlation coefficient.

A. SVR and the weight of the input variable

The SVR[8,9,10] for estimating the nonlinear function $f(\cdot)$ is

$$f(x_j) = w \bullet \phi(x_j) + b = \sum_{i=1}^{n} (\alpha_i - \alpha_i^*) K(x_i, x_j) + b \quad (5)$$

Where n is the length of the data; $\alpha_i - \alpha_i^*$ is support values and b is the bias.

Let $\phi(x)$ and w be, respectively, the nonlinear regressors and the associated weight. The kernel $K(x_i, x_j)$ is defined as a linear dot product of the nonlinear regressors, $\phi(x_i)$ and $\phi(x_j)$, given by

$$K(x_i, x_j) = \phi^T(x_i)\phi(x_j). \quad (6)$$

The associated weight is a p-dimensional vector,

$$w = (w_1, w_2, \cdots, w_p) = \sum_{i=1}^{n} (\alpha_i - \alpha_i^*) \phi(x_i), \quad (7)$$

Eq. (5) can be rewritten as

$$\begin{aligned} f(x_j) &= w \bullet \phi(x_j) + b \\ &= w_1 \cdot \phi(x_{j1}) + w_2 \cdot \phi(x_{j2}) + \cdots + w_p \phi(x_{jp}) + b \end{aligned} \quad (8)$$

Where $\phi(x_j) = \phi(x_{j1}), \cdots \phi(x_{jp})$ is a p-dimensional regressor.

If $w_m = 0, 1 \le m \le p$, it shows that the input variable m is irrelevant to the model output. So, the correlation between the input variable m and the model output can be quantified by the weight $|w_m|$ ($1 \le m \le p$).we can conclude that the smaller the weight $|w_m|$ is, the less important the input variable m is.

B. Weighted-KNN imputation method

According to the above analysis, the weights of SVR can be quantified by the correlation between the each input variable and the model output, which is given by (7).

After the weight is computed by the SVR method, the distance measurement between the missing data sample and the nearest neighbor sample can be modified as presented below. Rewriting (1), the distance can be expressed as

$$d_{ij} = \sqrt{\sum_{t=1}^{p} [w_t (x_{it} - x_{jt})]^2} \quad (9)$$

Let $x_i = (x_{i1}, \cdots, x_{ip})$ is a neighbor sample vector, $x_j = (x_{j1}, \cdots, x_{jp})$ is a missing data sample vector, $x_{it} (i = 1, \cdots, n; t = 1, \cdots, p)$ represents the value of variable t in sample x_i. w_t is the weight of SVR function and also is the input variable weight to the model output.

On the basis of the Weighted-KNN imputation method, the estimation of missing values can be obtained

$$x_{jt} = \sum_{i=1}^{k} q_{ij} x_{it}, (t = 1, \cdots, p) \quad (10)$$

Where $q_{ij} = \frac{s_{ij}}{\sum_{i=1}^{k} s_{ij}}$ and $s_{ij} = 1/d_{ij}$ be, respectively, the weight of neighbor sample corresponding to the missing sample and the distance similarity.

Here is step by step on how to compute the missing values based on the Weighted-KNN imputation algorithms:

Let $x_j = (x_{j1}, \cdots, x_{jp})$ denotes the missing data, $x_i = (x_{i1}, \cdots, x_{ip}), i = 1, \cdots, n$ indicates the complete data, where n is total number of complete samples, p is total number of feature variables.

Step 1) Find k similar samples $(x_1, \cdots, x_k)$ according to the distance measure between the missing data sample and the complete data samples.

Step 2)Estimate the missing dimension from remaining dimensions based on the SVR with k data samples.

Step 3) Compute weight vector

$$w = [w_1, w_2, \cdots, w_p] = \sum_{i=1}^{n} (\alpha_i - \alpha_i^*) \phi(x_i);$$

Step 4) Calculate the weighted distance between the k nearest neighbors and the missing sample x_j by

$$d_{ij} = \sqrt{\sum_{t=1}^{p} [w_t (x_{it} - x_{jt})]^2} \ ;$$

Step 5) Estimate the missing value:

$$x_{jt} = \sum_{i=1}^{k} q_{ij} x_{it}, (t = 1, \cdots, p)$$

Where $q_{ij} = \dfrac{s_{ij}}{\sum_{i=1}^{k} s_{ij}}$ and $s_{ij} = 1/d_{ij}$;

Step 6) Compute the missing variable value to impute the data set. Repeat Step5) until the all complete data are obtained.

IV. EXPERIMENT

In order to evaluate our method, the dataset [11] about corrosion behavior of 3C steel in different seawater environment was used as a source of data for this study. The original dataset consists of 46 samples, which is divided into complete dataset and missing dataset. According to the corrosion prior knowledge and the expert suggestion, five factors are identified. The seawater environmental factors important to corrosion rate of steels are followed by temperature, dissolved oxygen content, salinity, pH, oxidation-reduction potential (ORP). Here, T/ ℃ denotes temperature; DO/ mg • L^{-1}denotes dissolved oxygen content; Sal /ppt denotes salinity; ORP/V denotes oxidation-reduction potential; Rate/ μ A • cm^{-2} denotes corrosion rate (The text after "/"is unit). These data variables are frequently unavailable, it is important to impute them in predicting material corrosion in the process of materials selection in project designs.

The study idea is: A part of data variables among original corrosion data will be removed. Form the view of the relation of samples, the k samples similar to missing data are selected from the dataset by KNN method. Then, the k samples correspond to every missing data are used to establish the model based on the SVR for the purpose of quantifying the weight of the inputs corresponding to the outputs, on the basis of which, the weighted-KNN imputation method is to infer the values of missing variables removed.

In this case, we select the 4 data corresponding to 7,10,19,25 numbers as missing value dataset which was listed in Table 1(missing values in bold print), the remaining dataset consists of 42 samples to be used for the purpose of finding k samples similar to missing data. Distance measure results are reported in Table 1 illustrating that the k-nearest-neighbors most similar to every missing data sample. In this work, we tried several numbers and decided to use k=8 based on the accuracy of the model after the imputation process. Since we are exploiting correlations between different dimensions in our Weighted-KNN distance metric, we expect that we will achieve better performance on the dimensions that exhibit high correlations with each other. In this case, we can estimate the missing dimension from the remaining dimensions based on the SVR method.

Table 1. Missing value dataset

serial number	T (℃)	DO (mg • L^{-1})	Sal (ppt)	pH	ORP (mV)	Rate (μ A • cm^{-2})
7	27.87	**6.55**	31.68	7.2	**356.0**	14.06
10	29.37	6.82	**30.12**	**6.2**	414.0	17.11
19	**24.73**	**6.06**	17.33	7.88	321.0	11.446
25	25.57	**6.7**	32.19	**8.09**	325.0	11.872

Table 2. Similar Samples Selection

serial number of missing data	k-nearest-neighbors
7	3,5,6,8,13,14,31,43
10	2,5,9,15,16,17,26,36
19	11,18,20,21, 8,29,32,37
25	1,27,30,31,39,44,45,46

Then, the Weighted-KNN imputation method is used to complete the values of missing variables. The learning effect is compared among the four kernels: line kernel, polynomial kernel, Radial Basis Function kernel, Sigmoid kernel, the quadratic polynomial kernel is better than others to be determined as the kernel function of SVR imputation model. In order to evaluate our results, Root Mean Square error (*RMSE*), mean absolute error (*MAE*) and mean absolute percentage error (*MAPE*) were adopted for imputation performance evaluation. They are formulated by Eqs. (11), (12) and (13) respectively:

$$MAE = \frac{1}{m} \sum_{j=1}^{m} \left| \hat{y}_j - y_j \right|, \tag{11}$$

$$MAPE = \frac{1}{m} \sum_{j=1}^{m} \left| \frac{\hat{y}_j - y_j}{y_j} \right|, \tag{12}$$

$$RMSE = \frac{\sqrt{\sum_{j=1}^{m} (\hat{y}_j - y_j)^2}}{m-1}, \tag{13}$$

Where m denotes the number of test samples, y_j represents the jth target value, $\hat{y}_j$ stands for the predicted value for the jth test sample.

Table 2 and Table 3 display the results of Weighted-KNN imputation method and the KNN imputation method when missing variables are listed in the table respectively. From the results, it may be concluded that the proposed approach outperforms KNN imputation. Besides the above discussions, from Table 4, it can be found: the *MAE* of 4 test samples for Weighted-KNN is much smaller than that for KNN, which shows that the Weighted-KNN imputation performs significantly better than the KNN. Meantime, for Weighted-KNN, the *MAPE* of 4 test samples is also less than the KNN, which directly reflects the difference between the imputation value and the real value. The *RMS* of 4 test samples, 1.83 for Weighted-KNN and 4.47 for KNN, indicate that the imputation effect for the missing value by Weighted-KNN is better than that of KNN method.

Table 3. The estimation of missing data with Weighted-KNN

serial number	T (℃)	DO (mg • L^{-1})	Sal (ppt)	pH	ORP (mV)	Rate (μA • cm^{-2})
7	27.87	**7.13**	31.68	7.2	**362.0**	14.06
10	29.37	6.82	**36.47**	**7.09**	414.0	17.11
19	**25.59**	**5.38**	17.33	7.88	321.0	11.446
25	25.57	**7.26**	32.19	**9.21**	325.0	11.872

Table 4. The estimation of missing data with KNN

serial number	T (℃)	DO (mg • L^{-1})	Sal (ppt)	pH	ORP (mV)	Rate (μA • cm^{-2})
7	27.87	**8.35**	31.68	7.2	**373.0**	14.06
10	29.37	6.82	**39.59**	**8.53**	414.0	17.11
19	**26.33**	**8.42**	17.33	7.88	321.0	11.446
25	25.57	**9.75**	32.19	**11.07**	325.0	11.872

Table 5. Evaluation on the imputation results between WKNN and KNN

error	KNN	WKNN
MAE	11.815	4.26
MAPE	2.299	0.15
RMS	4.47	1.83

V. CONCLUSIONS

This paper studied a new imputation method towards the task of establishing a model from observation data when missing values occur among the multivariate input data. The main idea is to exploit correlations between different dimensions in Weighted-KNN distance metric when imputing the missing dimension, where each dimension should be weighted by the respective correlation coefficient obtained by the SVR method. The imputation method is stimulated by the steel corrosion dataset in seawater environmental, which was demonstrated to have superior results to the KNN imputation method.

ACKNOWLEDGMENT

This work is supported by Chinese national science and technology infrastructure platforms construction project (NO. 2005 DKA10400).

REFERENCES

[1] R.J.A. Little and D.B.Rubin, Statistical Analysis with Missing Data, New York, Wiley,1987.

[2] R.J.A. Little and D.B.Rubin, The Analysisy of Social Science Data with Missing Values, Sociological Methods and Research, vol.18, pp.292-326,1990.

[3] JinYongJin, Imputation adjustment method for missing data, Application of Statistics and Management, 2001, 20(5):47-53

[4] Imputation method for regional missing data using spatial auto-regression model, Application of Statistics and Management, 2005, 24(5):45-50

[5] HeKaiTao, ChenMing, ZhangZhiGuo. Complement of incomplete spatial information via RBF network[J], geological bulletin of china, 2005,24(5):476-479

[6] Jiawei Han, Micheline Kamber,Famin,MeingXiaofeng, Data Mining: Concepts and Techniques, Machinery Industry Press,2001.

[7] Jonsson Per, Wohlin Claes, An Evaluation of K-nearest Neighbours Imputation Using Likert data, Proceedings-10th International Symposium on Software Metrics, Metrics 2004, pp.108-118, 2004.

[8] Vapnik V N. The nature of statistical learning theory[M].NY: Springer- Verlag, 1995.

[9] Zhang Xue Gong, On the statistical learning theory and support vector machines[J], ACTA AUTOMATICA SINICA, 2000(1):32- 42.

[10] Vapnik V, Golowich S, Smola A.Support Vector method for function approximation, regression estimation, and signal processing[M]//Mozer M, Jordan M, Petsche T.Neural Information Processing Systems.[S.l.]:MIT Press, 1997.

[11] LiuXueQing, TangXiao, WangJia. Correlation between seawater environmental factors and marine corrosion rate using artificial neural network analysis, Journal of Chinese Society for Corrosion and Protection, 2005,25(1):11~14.