A SVM Regression Based Approach to Filling in Missing Values

Feng Honghai[1,2], Chen Guoshun[3], Yin Cheng[4],
Yang Bingru[2], and Chen Yumei[5]

[1] Urban & Rural Construction School, Hebei Agricultural University
071001 Baoding, China
honghf@mail.hebau.edu.cn
[2] Information Engineering School, University of Science and Technology Beijing
100083 Beijing, China
[3] Ordnance Technology Institute, Shijiazhuang
050000 Shijiazhuang, China
[4] Modern Educational Center, Hebei Agricultural University
071001 Baoding, China
[5] Tian'e Chemical Fiber Company of Hebei Baoding
071000 Baoding, China

Abstract. In KDD procedure, to fill in missing data typically requires a very large investment of time and energy - often 80% to 90% of a data analysis project is spent in making the data reliable enough so that the results can be trustful. In this paper, we propose a SVM regression based algorithm for filling in missing data, i.e. set the decision attribute (output attribute) as the condition attribute (input attribute) and the condition attribute as the decision attribute, then use SVM regression to predict the condition attribute values. SARS data set experimental results show that SVM regression method has the highest precision. The method with which the value of the example that has the minimum distance to the example with missing value will be taken to fill in the missing values takes the second place, and the mean and median methods have lower precision.

1 Introduction

Because of the "garbage in, garbage out" principle, data quality problems can be very expensive - "losing" customers, "misplacing" billions of dollars worth of equipment, misallocated resources due to glitches forecasts, and so on. Solving data quality problems typically requires a very large investment of time and energy - often 80% to 90% of a data analysis project is spent in making the data reliable enough so that the results can be trustful.

Data in the real world are often plagued by missing, ambiguous values that can greatly hinder some types of analysis. There are many schemes for guessing the identity of such values, for example, using logistic regression [1], by assuming that the missing points are the same as their nearest neighbors or the same as the most abundant data type within some radius.

Complete-case analysis [2], where cases with missing values are discarded, is often conducted because its simplicity and the comparability of univariate statistics. How-

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 581–587, 2005.

ever, discarding incomplete cases may lead to a considerable loss of information and, moreover, to serious biases in estimates [2]. Means and regression imputation [2,3,4] are widely used, due to their quickness and simplicity and lack of easy-to-use software packages that implement more advanced methods, such as EM imputation [2,3,4]. For example, numbers describing the central tendency (for example mode, median or mean) have often been used in machine learning studies to treat missing values.

Recently, K-Nearest Neighbor (KNN), sample mean imputation (SMI) [7-9], multivariate regression [10-12], mixture of principal component analyzers (MPCA) and variation Bayes (VB) etc data mining methods have been introduced to carry out the imputation of missing data.

In this paper we propose a SVM regression based algorithm to fill in missing data, i.e. set the decision attributes (output or classes) as the condition attributes (input attributes) and the condition attributes as the decision attributes, so we can use SVM regression to predict the missing condition attribute values. The SARS data experiments show that our methods are available.

2 Support Vector Machine [5]

Support Vector (SV) machines comprise a new class of learning algorithms, motivated by the results of the statistical learning theory. SV regression estimation seeks to estimate functions

$$f(\mathbf{x}) = (\mathbf{w} \cdot \mathbf{x}) + b, \qquad \mathbf{w}, \mathbf{x} \in \mathbb{R}^N, \qquad b \in \mathbb{R} \tag{1}$$

based on data

$$(\mathbf{x}_1, y_1), \cdots, (\mathbf{x}_l, y_l) \in \mathbb{R}^N \times \mathbb{R}, \tag{2}$$

by minimizing the regularized risk functional

$$\|\mathbf{W}\|^2 / 2 + C \bullet R_{\text{emp}}^{\varepsilon}. \tag{3}$$

where C is a constant determining the trade-off between minimizing the training error, or empirical risk

$$R_{\text{emp}}^{\varepsilon} = \frac{1}{l} \sum_{i=1}^{l} \left| y_i - f(\mathbf{x}_i) \right|_{\varepsilon}$$

and the model complexity term $\|\mathbf{W}\|^2$. Here, we use the so-called ε -insensitive loss function

$$\left| y - f(\mathbf{x}) \right|_{\varepsilon} = \max \left\{ 0, \left| y - f(\mathbf{x}) \right| - \varepsilon \right\}$$

The main insight of the statistical learning theory is that in order to obtain a small risk, one needs to control both training error and model complexity, i.e. explain the

data with a simple model. The minimization of Eq. (3) is equivalent to the following constrained optimization problem (Vapnik, 1995):
minimize

$$\tau(\mathbf{w},\xi^{(*)}) = \frac{1}{2}\|\mathbf{w}\|^2 + C\frac{1}{l}\sum_{i=1}^{l}(\xi_i + \xi_i^*) \tag{4}$$

subject to the following constraints

$$((\mathbf{w}\bullet\mathbf{x}_i)+b) - y_i \le \varepsilon + \xi_i \tag{5}$$

$$y_i - ((\mathbf{w}\bullet\mathbf{x}_i)+b) \le \varepsilon + \xi_i^* \tag{6}$$

$$\xi_i^{(*)} \ge 0, \qquad \varepsilon \ge 0 \tag{7}$$

As mentioned above, at each point $\mathbf{x}_i$ we allow an error of magnitude ε Errors above ε are captured by the slack variables ξ^* (see constraints (5) and (6)). They are penalized in the objective function via the regularization parameter C chosen a priori (Vapnik, 1995).

In the ν -SVM the size of ε is not defined a priori but is itself a variable. Its value is traded off against model complexity and slack variables via a constant $\nu \in (0,1]$
minimize

$$\tau(\mathbf{W},\xi^{(*)},\varepsilon) = \frac{1}{2}\|\mathbf{W}\|^2 + C\bullet(\nu\varepsilon + \frac{1}{l}\sum_{i=1}^{l}(\xi_i + \xi_i^*)) \tag{8}$$

subject to the constraints (5)–(7). Using Lagrange multipliers techniques, one can show (Vapnik, 1995) that the minimization of Eq. (4) under the constraints (5)–(7) results in a convex optimization problem with a global minimum. The same is true for the optimization problem (8) under the constraints (5)–(7). At the optimum, the regression estimate can be shown to take the form

$$f(\mathbf{x}) = \sum_{i=1}^{\ell}(\alpha_i^* - \alpha_i)(\mathbf{x}_i\bullet\mathbf{x}) + b. \tag{9}$$

In most cases, only a subset of the coefficients $(\alpha_i^* - \alpha_i)$ will be nonzero. The corresponding examples $\mathbf{x}_i$ are termed support vectors (SVs). The coefficients and the SVs, as well as the offset b ; are computed by the ν -SVM algorithm. In order to move from linear (as in Eq. (9)) to nonlinear functions the following generalization can be done: we map the input vectors $\mathbf{x}_i$ into a high-dimensional feature space Z through some nonlinear mapping, $\Phi : \mathbf{X}_i \to \mathbf{Z}_i$ chosen a priori. We then solve the optimization problem (8) in the feature space Z . In that case, the inner product of the input vectors $(\mathbf{x}_i\bullet\mathbf{x})$ in Eq. (9) is replaced by the inner product of their icons in

feature space Z, $(\Phi(\mathbf{x}_i) \bullet \Phi(\mathbf{x}))$ The calculation of the inner product in a high-dimensional space is computationally very expensive. Nevertheless, under general conditions (see Vapnik, 1995 and references therein) these expensive calculations can be reduced significantly by using a suitable function k such that

$$(\Phi(\mathbf{x}_i) \bullet \Phi(\mathbf{x})) = k(\mathbf{x}_i \bullet \mathbf{x}), \tag{10}$$

leading to nonlinear regression functions of the form:

$$f(\mathbf{x}) = \sum_{i=1}^{\ell} (\alpha_i^* - \alpha_i) k(\mathbf{x}_i, \mathbf{x}) + b \tag{11}$$

The nonlinear function k is called a kernel (Vapnik, 1995). In our work we use a Gaussian kernel

$$k(\mathbf{x}, \mathbf{y}) = \exp(-\left\|\mathbf{x} - \mathbf{y}\right\|^2 / (2\sigma_{\text{kemel}}^2)) \tag{12}$$

3 Algorithm for Filling Missing Data

(1) Select the examples in which there are any not missing attribute values.

(2) Set one of condition attributes (input attribute), some of whose values are missing, as the decision attribute (output attribute), and the decision attributes as the condition attributes by contraries.

(3) Use SVM regression to predict the decision attribute values.

4 Experiment and Results

The experiments are done using the LIBSVM [6] software package on SARS data. The SARS data are obtained from the analysis of microelements Zn Cu Fe Ca Mg K Na in one's body. The category or class labels are 1 and 0, where 1 denotes that the patients are infected by SARS, and 0 not infected. Some examples of the whole data set are in Table 1 and the experiment results are given in Table 2, Table 3, Table 4.

Table 1. Some examples of whole SARS data set

Class	Zn	Cu	Fe	Ca	Mg	K	Na	Class	Zn	Cu	Fe	Ca	Mg	K	Na
1	164	22.2	35.5	2212	281	153	549	0	166	15.8	24.5	700	112	179	513
1	173	8.99	36.0	1624	216	103	257	0	185	15.7	31.5	701	125	184	427
1	202	18.6	17.7	3785	225	31.0	67.3	0	193	9.80	25.9	541	163	128	642
1	182	17.3	24.8	3073	246	50.7	109	0	159	14.2	39.7	896	99.2	239	726
1	211	24.0	17.0	3836	428	73.5	351	0	226	16.2	23.8	606	152	70.3	218
1	246	21.5	93.2	2112	354	71.7	195	0	171	9.29	9.29	307	187	45.5	257
1	164	16.1	38.0	2135	152	64.3	240	0	201	13.3	26.6	551	101	49.4	141

Table 2. Experiment results of filling in attribute Ca's values

Supposing that attribute Ca's values are missing								results of guessing Ca's values based on SVM regression							
Ca	Zn	Cu	Fe	Class	Mg	K	Na	Ca	Zn	Cu	Fe	Class	Mg	K	Na
2157	209	6.43	86.9	1	288	74.0	219.8	2158.23	209	6.43	86.9	1	288	74.0	219.8
3870	182	6.49	61.7	1	432	143	367.5	3869.39	182	6.49	61.7	1	432	143	367.5
1806	235	15.6	23.4	1	66	68.9	188	1483.64	235	15.6	23.4	1	166	68.9	188

Table 3. Experiment results of filling attribute Mg's values

Supposing that attribute Mg's values are missing								results of guessing Mg's values based on SVM regression							
Mg	Ca	Zn	Cu	Fe	Class	K	Na	Mg	Ca	Zn	Cu	Fe	Class	K	Na
288	2157	209	6.43	86.9	1	74.0	219.8	287.17	2157	209	6.43	86.9	1	74.0	219.8
432	3870	182	6.49	61.7	1	143	367.5	431.80	3870	182	6.49	61.7	1	143	367.5
166	1806	235	15.6	23.4	1	68.9	188	166.10	1806	235	15.6	23.4	1	68.9	188

In Table 1, attribute "class" is the output attribute or decision attribute, "1" denotes the patient suffers from SARS. We can use standard SVM to estimate a new example's class which it belongs to.

However, if there are some missing values in an input attribute (condition attribute), the SVM method cannot be used directly. So we set the input attribute as the output attribute or decision attribute, and set the original output attribute as one of the input attributes. Finally, use SVM regression to predict the missing values.

Table 4. Comparative results of filling in attribute Mg's and Ca's values with different methods

Real values		Method (1)		Method (2)		Method (3)		Method (4)		Method (5)	
Mg	Ca	Mg	Ca	Mg	Ca	Mg	Ca	Mg	Ca	Mg	Ca
288	2157	287.17	2158.23	113.4	2511.1	108	2220	215.5	1882.5	354	2112
432	3870	431.80	3869.39	113.4	2511.1	108	2220	202.8	1546.8	428	3836
166	1806	166.10	1483.64	113.4	2511.1	108	2220	209.1	1714.6	216	1624

In Table 2, the original output attribute is the "class". If we suppose that attribute Ca's values are missing, the attribute "class" should be set as one of the input attributes, and the attribute "Ca" be set to be the output attribute, so we can use the SVM regression method to predict the missing values of attribute Ca.

In Table 4, the left two columns are real values. Method (1) denotes the SVM regression methods proposed in this paper. In Method (2), the mean of all the values of the same class will be taken to fill in the missing values. In Method (3), the median of all the values of the same class will be taken to fill in the missing values. In Method (4), the mean of the two closest neighbor values (natural order) will be taken to fill in the missing values. In Method (5), for the example with the missing value we select the example that has the minimum distance to it, and take the value of the same attribute which the missing value belongs to to fill the missing value, i.e., value of the example that has the minimum distance to the example that contains the missing value will be taken to fill in the missing value.

In Table 4, obviously, the SVM regression method has the highest precision, Method (5) takes second place in precision, and the other methods have lower precision.

5 Discuss and Future Works

(1) The experimental results indicate that the SVM regression based algorithm for filling in missing data is available.

(2) Since the support vectors influence greatly the results of regression, the training data set had best be selected to be complete, i.e., we should select enough complete examples where there are not missing data as the training data set. If there are not enough complete examples in the training data set the regression accuracy will be influenced.

(3) The regression methods give a comprehensive and average guess for the missing data, the data, which have been filled in, reflect or embody the holistic information hidden in the whole data set, and the local information may be ignored or be submerged. This is in contrast to methods such as by assuming that the missing points are the same as their nearest neighbors where the local information is taken into account, and the holistic information ignored, resulting in bigger errors.

(4) Our future works will be the followings: (1) comparative research on different algorithms for filling in missing data such as EM, ANN etc. (2) implement the experiment on large data set.

References

1. T.M. Thomas, K.R. Plymat, J. Blannin, T.W. Meade: Prevalence of Urinary Incontinence, Br. Med. J. 281 (1980) 1243-1245.
2. R.J.A. Little, D.B. Rubin: Statistical Analysis with Missing Data, Wiley, New York, (1987).
3. J.L Schafer: Analysis of Incomplete Multivariate Data, Chapman & Hall, London, (1997).
4. M.A. Hill: SPSS Missing Value Analysis 7.5, SPSS Inc., Chicago, (1997).
5. Vapnik V N: The Nature of Statistical Learning Theory. NY: Springer-Verlag, (1995).
6. Chang. C. & Lin, C: (2001). LIBSVM: a library for support vector machines. Software is available for download at http://www.csie.ntu.edu.tw/~cjlin/libsvm
7. Zhao Guanghui, Song Huazhu, Xia Hongxia, Zhong Luo: Comparison of Missing Data Estimation Methods in Satellite Information for Scientific Exploration. DCABES (2004) 278-280
8. M. H. Cartwright, M. J. Shepperd, and Q. Song: Dealing with Missing Software Project Data. 9th International Software Metrics Symposium. (2003) 154-165
9. Eduardo R. Hruschka, Estevam R. Hruschka and Nelson F. F. Ebecken: Evaluating a Nearest-Neighbor Method to Substitute Continuous Missing Values. Lecture notes in computer science (2003) 723-734
10. Th. Liehr: data Preparation in Large Real-World data Mining Projects: Methods for Imputing Missing Values. Exploratory data analysis in empirical research (2003) 248-256

11. Jau-Ji Shen; Ming-Tsung Chen: A Recycle Technique of Association Rule for Missing Value Completion. 17th International Conference on Advanced Information Networking and Applications. (2003) 526-529
12. Mehtap KANDARA, Osman KANDARA: Association Rules to Recover the Missing Data Value for An Attribute in a Database. The 7th World Multiconference on Systemics, Cybernetics and Informatics (2003) 1-6
13. Shigcyuki. Oba, Masa-aki. Sato, Ichiro. Takemasa, Morito. Monden, Ken-ichi. Matsubara and Shin Ishii: Missing Value Estimation Using Mixture of PCAs. International Conference on Artificial Neural Networks. (2002) 492-497
14. Jerzy W. Grzymala-Busse, Ming Hu: A Comparison of Several Approaches to Missing Attribute Values in Data Mining. 2nd International Conference on Rough Sets and Current Trends in Computing (2000) 378-385