

A New Imputation Method based on GMDH

He Changzheng, Zhu Bing

Business School Sichuan University, Chengdu 610064, P.R. China

Abstract

Most existing imputation methods do not take noise into consideration, which is rarely the case in reality. In this paper, we combine Group Method of Data Handling (GMDH) and the well-known Expectation Maximization (EM) algorithm and propose a new imputation algorithm to deal with missing values in noisy data. Numerous experiments and comparative studies on four UCI datasets show that our method GMDH imputation is more robust to noise than the other imputation methods used as benchmark at high noise level.

Keywords: Missing values, Noise, imputation, GMDH

1. Introduction

Real-world data in economy and business is often corrupted by missing values, especially the data collected from surveys. For example, consumer data obtained from questionnaires usually contains missing data because the consumers refuse to answer some sensitive questions (e.g., income level, age) or they just have no opinions about them and so on. This missingness complicates the data mining process because most data mining algorithms cannot be immediately and straightforwardly applied to data with missing values. In order to solve this problem, two categories of techniques have been developed. First, there are missing data toleration techniques which integrate the techniques of missing values handling in specific data mining algorithm such as classification[1-3], clustering[4] and feature selection[5]. Second, there are missing data imputation techniques which fill in missing values before using complete-data method. One advantage of imputation is that it is independent of learning algorithm. This allows the user to select a suitable learning algorithm after imputation. Therefore, imputation has received considerable attention and a great number of methods have been proposed in recent years [6, 7].

However, an important issue has been neglected by previous research: missing values do not stand alone, and they are usually accompanied by noise which comes from the process of data collection, data entry and data transformation, etc. The existence of noise may introduce some negative effects. For example, the noise can reduce the accuracy of classifiers [8]. Consequently, it will affect the performance of imputation methods that are based on these classification algorithms. Therefore, although many imputation methods are effective in noise-free environment, they may perform poorly in noisy datasets.

In this paper, we provide an extensive evaluation of the impacts of noise on some popular imputation methods and present a new robust method to impute missing values in noisy environment. The new method integrates EM algorithm with Group Method of Data Handling (GMDH) which is proposed by Ivakhnenko [9]. This integration is motivated by two observations. One is that GMDH has been recognized as a noise-immunity approach[10] and successfully been applied to many real-world data mining applications in which data are noisy[11]. The other observation is that EM algorithm is a useful framework in estimating missing data [12, 13]. Experiments on four UCI datasets from different domains show that our method is more robust to noise than the other imputation methods used as benchmark at high noise level.

2. GMDH Imputation

The main idea of GMDH Imputation is that: Integrating GMDH into the framework of the EM algorithm and expecting it will have an advantage over traditional EM imputation and other approaches in noise-immunity. Let D denote an incomplete dataset with r variables $D = \{A_1, A_2, \dots, A_r\}$ and n instances. For each variable A_j , $j = 1, 2, \dots, r$. It contains two parts: $A_j = \{A_j^{obs}, A_j^{mis}\}$, where A_j^{mis} are missing elements in A_j and A_j^{obs} are observed elements in A_j . Similarly, the entire dataset D also consists of two components, $D = \{D^{obs}, D^{mis}\}$, where D^{obs} is the set of observed values of D and D^{mis} is the set of missing values. The aim of imputation is to fill in all the blanks of incomplete

dataset D , so that the estimated complete dataset \hat{D} can be used for succeeding data mining algorithm. A good imputation method will make the imputed value \hat{D}^{mis} approximate the true of the missing elements as close as possible.

The GMDH Imputation first fills in the original incomplete dataset D by simple mean/mode imputation (see Step 1) to get an initial complete dataset D' . Then it updates these initial estimated missing values variable by variable using the iterative process of EM algorithm as follows. Given a variable A_j , we first build a model f_j with GMDH by treating f_j as dependent variable and other variables as independent variables in the E-step (Step 2). Then the missing elements A_j^{mis} at variable A_j are replaced by its estimates from model f_j in the M-step (Step 3). This process of model generating and missing value replacing is repeated iteratively until the estimates of missing elements do not change or maximal number of iterations is reached. The main steps of the GMDH Imputation are shown in Algorithm GMDH Imputation

Algorithm GMDH Imputation

Input:

D - a $n \times r$ incomplete dataset

Output:

\hat{D} - a $n \times r$ complete dataset

- Step 1:** For each variable A_j in D , replace missing elements A_j^{mis} by mean (if A_j is a numeric variable) or mode (if A_j is a nominal variable) of the observed elements A_j^{obs} to get initial complete dataset D' and let $\hat{D} = D'$;
- Step 2:** Use A_j as dependent variable ($y = A_j$) and all the remaining variables as independent variables ($\mathbf{x} = \{A_s \mid s = 1, \dots, r, s \neq j\}$) to build model $y = f_j(x)$ by GMDH;
- Step 3:** Obtain the estimates \hat{A}_j of the variable A_j from model f_j , $\hat{A}_j = f_j(\mathbf{x})$, and then use the estimates of missing elements \hat{A}_j^{mis} to update the missing elements A_j^{mis} , $A_j^{mis} = \hat{A}_j^{mis}$;
- Step 4:** Repeat Step 3-4 until maximum number of iterations is reached or the estimates of missing elements \hat{A}_j^{mis} cease to change much. Then assign current values of A_j to the corresponding elements in \hat{D}
- Step 5:** Following Step 2-4 to updating the missing values of remaining variables

3. Experiments

3.1 Experimental setting

In this section, we will evaluate the robustness of GMDH Imputation in noisy environment through experiments. Four datasets from the UCI ML repository [14] are used in the experiments. The basic information of these datasets is listed in Table 1. The four datasets are chosen because they have no missing data (for the Breast dataset 16 instances with missing values are removed). Consequently, we can have total control over the generation of missing data in the dataset to produce missing data with specified pattern and evaluate the performance of imputation methods by comparing the imputed values with original ones.

Table 1 Datasets used in the experiments

Dataset	Size	#Attr
Balance	625	5
Breast	683	10
Cmc	1473	10
Iris	150	5

To verify the effectiveness of GMDH Imputation, four popular imputation methods are used in our experiments as base line. They are regression imputation (RI), EM imputation (EI), grey-based nearest neighbor imputation (GBNN), multiple imputation based on fully conditional specification (MI). Our experiments consider the following three aspects:

•**Missing rate:** 5%, 10%, and 20%

•**Noise level:** 20%

•**Missingness mechanisms:** MCAR, MAR, NMAR

In total, $3 \times 3 = 9$ scenarios (combination of missing rate, noise level, and missingness mechanism) are considered for every dataset in our experiments, we assume that all the variables have the same level of noise and the all the variables with missing values have the same missing rate and missing mechanism. To avoid bias, five independent experiments (runs) are implemented for each scenario of the dataset.

To evaluate the precision of imputation, the normalized mean absolute error (NMAE) is used and its value at variable A_j is calculated as follows:

$$NMAE_j = \begin{cases} \frac{1}{n_j^{mis}} \sum_{i=1}^{n_j^{mis}} \left(\frac{\hat{a}_{ij} - a_{ij}}{a_j^{\max} - a_j^{\min}} \right) & \text{if } A_j \text{ is numerical} \\ 1 - \frac{n_j^{cor}}{n_j^{mis}} & \text{if } A_j \text{ is nominal} \end{cases} \quad (1)$$

where n_j^{mis} is the number of missing values at A_j , a_{ij} and \hat{a}_{ij} denote the true value and imputed value of the missing data respectively, a_j^{\max} and a_j^{\min} are the maximum and minimum value at A_j , n_j^{cor} is the number of missing values that are correctly predicted. The NMAE on the whole dataset takes the average over all the variables. The NMAE of one imputation method at a scenario is calculated as the average over all the five run of that scenario.

3.2 Experimental results and analysis

To validate the robustness of GMDH Imputation in noisy environment, ANOVA model is used to analyze the experimental results. Analysis of variance (ANOVA) with unreplicated measure is used to analyze the experimental results on each dataset. ANOVA is statistical model that can be used to test the hypothesis that the each level of factors have equal means when there are many factors influencing the experimental results simultaneously [15]. Two popular post hoc tests are used for multiple comparison: Fisher's Least Significant Difference test (LSD) and Tukey's Honestly Significant Difference test (HSD). Table 2-Table 5 report the results of post-hoc analysis, where the first column gives the marginal mean of the five methods, and the second and third column provide the grouping the five methods according to LSD and HSD at 5% significance level. A, B and C in the tables denote the first, second and third group, respectively.

As can be seen from Table 2-Table 5, in terms of imputation error, GMDH Imputation gives the lowest error on all the four datasets. The next is EI and it takes the second place. According to the grouping of LSD and HSD test, GMDH Imputation belongs to the first group on all the four datasets and takes this position alone on two datasets with a probability of 2/4. EI belongs to the first group on one dataset but does not take it by itself. It also falls into the second group on three datasets

All the above results from Table 2-Table 5 have demonstrated that GMDH imputation achieves higher imputation accuracy at high noise level (20%) when comparing with other benchmark methods.

Table 2 Main effect of imputation method on dataset Balance

Method	Means	LSD	HSD
GMDH	0.2403	A	A
EI	0.2516	B	B
RI	0.2541	B	B
MI	0.2612	BC	B
GBNN	0.2670	C	C

Table 3 Main effect of imputation method on dataset Breast

Method	Means	LSD	HSD
GMDH	0.1002	A	A
EI	0.1071	B	B
GBNN	0.1084	B	B
RI	0.1171	C	C
MI	0.1189	C	C

Table 4 Main effect of imputation method on dataset Cmc

Method	Means	LSD	HSD
GMDH	0.2314	A	A
EI	0.2350	A	A
RI	0.2364	A	A
GBNN	0.2376	A	A
MI	0.2532	B	B

Table 5 Main effect of imputation method on dataset Iris

Method	Means	LSD	HSD
GMDH	0.0963	A	A
GBNN	0.1003	A	A
MI	0.1168	AB	B
EI	0.1296	BC	B
RI	0.1510	C	C

4. Conclusion

In this paper, we have designed a robust method GMDH imputation which combines GMDH and EM algorithm together to impute missing values in noisy environment. Comparative studies have shown that GMDH Imputation performs quite well in comparison with other four popular imputation methods at high noise level. Given the frequent occurrence of missing values and noise, GMDH Imputation is a good choice in imputing incomplete data and has great potential in the real-world data mining applications.

References

- Williams, D., et al., *On classification with incomplete data*. Ieee Transactions on Pattern Analysis and Machine Intelligence, 2007. **29**(3): p. 427-436.
- Lim, C.P., J.H. Leong, and M.M. Kuan, *A hybrid neural network system for pattern classification tasks with missing features*. Ieee Transactions on Pattern Analysis and Machine Intelligence, 2005. **27**(4): p. 648-653.
- Saar-Tsechansky, M. and F. Provost, *Handling missing values when applying classification models*. Journal of Machine Learning Research, 2007. **8**: p. 1625-1657.
- Hathaway, R.J. and J.C. Bezdek, *Clustering incomplete relational data using the non-Euclidean relational fuzzy c-means algorithm*. Pattern Recognition Letters, 2002. **23**(1-3): p. 151-160.
- Aussem, A. and S.R.d. Morais. *A Conservative Feature Subset Selection Algorithm with Missing Data*. in *Data Mining, 2008. ICDM '08. Eighth IEEE International Conference on*. 2008.
- Little, R.J.A. and D.B. Rubin, *Statistical analysis with missing data*. 2002, New York: Wiley.
- Tsikriktsis, N., *A review of techniques for treating missing data in OM survey research*. Journal of Operations Management, 2005. **24**(1): p. 53-62.
- Zhu, X. and X. Wu, *Class Noise vs. Attribute Noise: A Quantitative Study*. Artificial Intelligence Review, 2004. **22**(3): p. 177-210.
- Ivakhnenko, A.G., *The Group Method of Data Handling in Prediction Problems*. Soviet Automatic Control, 1976. **9**(6): p. 21-30.
- Ivakhnenko, A. and V. Stepashko, *Noise Stability of Modeling*. 1985, Kiev: Naukova Dumka.
- Stepashko, V.S., *Noise-immunity of model selection based on prediction balance criterion*. Automatics, 1984(5).
- Nijman, M.J. and H.J. Kappen, *Symmetry breaking and training from incomplete data with Radial Basis Boltzmann Machines*. Int J Neural Syst, 1997. **8**(3): p. 301-15.
- Ghahramani, Z., et al., *Supervised learning from incomplete data via an EM approach*. Advances in Neural Information Processing Systems, 1994. **6**: p. 120-127.
- Blake, C.L. and C.J. Merz, *UCI repository of machine learning databases*. 1998.
- Miller, R., *Beyond ANOVA: basics of applied statistics*. 1997, Boca Raton, FL: Chapman & Hall.