Cost-Sensitive Imputing Missing Values with Ordering

Xiaofeng Zhu[1], Shichao Zhang[1,*], Jilian Zhang[1], Chengqi Zhang[2]

[1] Department of Computer Science, Guangxi Normal University, China
zhu0011@21cn.com; zhangsc@mailbox.gxnu.edu.cn; zhangjilian@yeah.net
[2] Faculty of Information Technology, University of Technology Sydney, Australia.
chengqi@it.uts.edu.au

Abstract

Various approaches for dealing with missing data have been developed so far. In this paper, two strategies are proposed for cost-sensitive iterative imputing missing values with optimal ordering. Experimental results demonstrate that proposed strategies outperform the existing methods in terms of imputation cost and accuracy.

Introduction

Missing value is an unavoidable problem, and various approaches for dealing with missing data have been developed. In fact, it is very important to consider the imputation ordering (ordering means which missing value should be imputed at first with the help of a specific criterion) during the imputation process, because not all attributes have the same impact on the imputation results. Usually, the higher correlation between the non-target attributes and the target attributes, the more important the attribute is. On the other hand, imputation ordering is important for reducing costs when we impute a missing value involving costs. However, to our knowledge, there are only few reports on improving the classification accuracy by ordering, for example, Claudio (2003), Numao (1999), and Estevam (2006). In this paper we present two strategies with imputation ordering to minimize imputation cost and improve the accuracy. One is called incremental iterative method, in which each last imputed information are added to training set for imputing the remained missing values, and it is repeated until the accuracy doesn't increase again. The other is the iterative method, in which each missing value is imputed with all information of the dataset including the instances with missing values that are first imputed by mean/mode method. Then it repeats until the accuracy doesn't increase.

Methodology

1) Missing Values Ranking

At first, Economic Criterion (EC) is applied to evaluate each attribute for finding the most economical attribute. EC is defined based on a relationship "benefit/cost". i.e., EC= cost/MI. Where *'cost'* is the sum of all cost for the attribute, including imputation cost and misclassification cost, in this paper, and we regard all the costs as a unit such as dollar. *'MI'* means mutual information. Attributes that have low MI with respect to the class label have less chance to participate in the final policy, so that properly filling the missing values for such attributes will have very low impact on the accuracy of the final policy, vice versa.

For each missing value, an imputation model will be built with the best use of all the observed information in order to get the optimal system performance; the more information for each missing value that can be observed, the more confident are the imputation results. The term Efficient Information (EI) is defined as the percentage of all instances that can be used for constructing imputation model in the dataset. In this paper, the efficient information (EI) for the missing value in i-th instance and j-th attribute is defined as follows:

$$EI_{i,j} = \frac{\textit{Useful}\ \text{Information for the Mising Vlaues}}{\text{All Information in the Dataset}}$$

Obviously, the more value of EI, the better performance the system is.

At last, missing values ranking will be considered both EC and EI by employing harmonic mean that allows us to specify the desired trade-off between EC and EI through a coefficient $\alpha \in (0, +\infty)$ (in our experiments, let $\alpha = 0.5$). Assuming EC_j has a weight of 1, and $EI_{i,j}$ has a weight of α.

$$Rank(i,j) = \frac{(\alpha+1)EI_{i,j} * EC_j}{EI_{i,j} + \alpha EC_j}$$

Where the number of instances is i, and j is the number of attributes. Computing $Rank(i,j)$ for each missing value, sorting them in descend ordering. Any imputation method can be applied in our two strategies, in this paper we use the non-parametric imputation method, such as, C4.5 algorithm, non-parametric kernel imputation method.

Note that, MVS: the set of missing values; OS: the set of observed instances; CIV: current imputed missing value which contain the maximum of $Rank(i,j)$; CS: completed information of CIV;

2) Incremental Imputation Strategy

logo1: For each iterative imputation
Do

*Corresponding author: Shichao Zhang

```
        Building a classifier for CIV based on CS;
        Imputing CIV based on this classifier;
         MVS= MVS -CIV;
        OS= OS+{CIV};
  While{MVS=∅ } ;  //finished all missing values
  Computing  CA(classification accuracy);
  Comparing current CA with the last CA
  If (current CA> last CA)
     IT++;          //IT: Imputation Times
     Goto  logo1:  //continue the other iterative imputation
  Else
      End the imputation;
      IT-1 is the imputation times of the algorithm;
```

3) Iterative Imputation Strategy

F*or the first imputation*

Each missing value imputed as

Mean for continuous attribute;

Mode for discrete attribute;

Since the second imputation

Method same as the Incremental Imputation Strategy;

Note that, there are two differences between the incremental algorithm and the iterative ones. At first, the former method constructs the complete information for CIV with all the observed information in the dataset. However, all the instances in the dataset are regarded as the complete information in iterative algorithm. Secondly, IT-1 in former method is the last imputation times. However, IT is the imputation times of the iterative algorithm because the last method has finished a first imputation.

Experiments and Results

Four UCI datasets are applied to our experiments, i.e. "Abalone", "Ecoli", "Pima" and "Vowel". There are not missing in these four datasets and the conditional attributes values are missed at random with missing rate 10%, 20% and 30%.

At first, we compare our two methods with a lexicographic ordering method in Claudio (2003) about the iterative imputation times for these datasets with missing. Our two methods have significant results than the lexicographic algorithm especially in the moderate missing rate such as 20%, our iterative method end the repeat till 8th iterative imputation, the incremental method's is 10, and the lexicographic method's is 18.

Then, we compare these three methods with un-ordering method that does not consider the imputation ordering by $MSE = m^{-2} sqrt(\sum_{i=1}^{m}(e_i - \tilde{e}_i)^2)$ in dataset "Abalone" whose conditional attributes are real values. Where e_i is the original attribute value; $\tilde{e}_i$ is the estimated attribute value, and m is the total number of missing. As well, we compute the correlation coefficient between the actual and predicted values. Results appear as follow.

	10%		20%		30%	
	MSE	ρ	MSE	ρ	MSE	ρ
Incremental	618.3	0.98	692.4	0.96	847.9	0.92
Iterative	627.4	0.98	698.1	0.95	856.3	0.91
lexicographic	785.1	0.95	805.6	0.90	902.4	0.89
Un-order	792.6	0.91	884.3	0.87	956.7	0.86

These results show that the ordering methods completely dominate the un-ordering method and that accounting for the ordering method imputes each missing with the best efficiency.

At last, we design some experiments to demonstrate both the classification accuracy and cost assuming each attribute has a cost. The experimental results test all the ordering imputation methods can get better classification accuracy in same fixture cost, as well as the ordering methods paid less cost in order to get the fixture classification accuracy. On the other hand, our two methods are better than the lexicographic method.

Ongoing and Future Work

Subsequent to the preliminary results outline above, we are actively working on analyzing further and comparing the efficiency of our two algorithms. We plan to assess the two algorithms with different classifiers or different ordering conditions. In our experiments, we found the different α values have some impact for the performance of the algorithm. We then plan to explore the advantages and disadvantages on the different coefficient α both in experiments and in theory.

Acknowledgement

This work is partially supported by Australian large ARC grants (DP0449535, DP0559536 and DP0667060), a China NSF major research Program (60496327), a China NSF grant (60463003), a National Basic Research Program of China (2004CB318103), an Overseas Outstanding Talent Research Program of Chinese Academy of Sciences (06S3011S01), and an Overseas-Returning High-level Talent Research Program of China Human-Resource Ministry, and Innovation Plan of Guangxi Graduate Education (2006106020812M35).

References

Claudio (2003): Incremental Tree-Based Missing Data Imputation with Lexicographic Ordering. *Computing Science and Statistics*, 35, 2003.

NUMAO et al. (1999): Ordered estimation of missing values. *PAKDD99*, pp 499-503.

Estevam et al. (2006): Bayesian network for imputation in classification problems. *Journal of Intelligence Information System*, DOI 10.1007/s 10844 –006 –0016 -x.