A Quantitative Study of the Effect of Missing Data in Classifiers

Peng Liu, Lei Lei, Naijun Wu
School of Information Management and Engineering,
Shanghai University of Finance & Economics, Shanghai, 200433, China
liupeng@mail.shufe.edu.com

Abstract

In data mining approaches, predictive classification has a wide range of application. However, there are always missing data in the datasets, which affect the accuracy of classifiers. This paper will investigate the influence of missing data to classifier. The sensitivity analysis of six classifiers to missing data is studied in experiments. The results indicate that, in the datasets, when the proportion of missing data exceeds 20%, they do have a huge adverse effect on the prediction accuracy. Among the six classifiers, the Naive Bayesian classifier is the least sensitive to missing data. For the popular missing data treatment methods using prediction model to handle missing data, Naive Bayesian classifier will be preferred.

1. Introduction

Data Mining (DM) is the process of discovering interesting knowledge from large amounts of data stored either in databases, data warehouse, or other information repositories [1]. There are a lot of functions of DM, such as description, association analysis, classification and prediction, clustering analysis etc. Among all, classification and prediction are widely used in many fields. However, in real-world datasets, there are many problems in data quality such as incompleteness, redundancy, inconsistency, noise data etc. All these serious data quality problems affect the performance of DM algorithms [2]. Missing data is a quite common problem of data quality in real-life datasets. Then, what is the reaction of classifier to missing data?

For the large amount of missing data in real-word datasets, several methods have been proposed. For example, case deletion, mean imputation and model prediction. The basic idea of model prediction is using prediction model, built on known data, to predict and fill in the missing data. The performance of the missing treatment relies on the prediction model. Then, which classifier will be suitable for missing data handling?

This paper will emphasize missing data’s impact on classifiers. Basic knowledge about classifier and treatment methods for missing data which using classifier is introduced in the next section. In Section 3, sensitivity analysis is applied to missing data in classifiers in our experiments and the effects of missing data are evaluated.

2. Classification and Prediction

2.1. Classifier

Classification means constructing a classifying function or model from the known data. Such
function or model can also be called ‘classifier’, which can classify the records in database into given classes, thus can predict the unknown variables under some given conditions [3].

Classifiers differ greatly in prediction accuracy, training time and number of leaves (Decision Trees). There is not a classifier which performs best in all aspects [4]. Prediction accuracy of classifiers can be affected by the factors as follows [3].

**Number of records in training subset.** Classifier needs to learn from training set. Therefore, larger training set makes the classifier more reliable. But the training time is also become longer.

**Data quality.** Problems such as noise data, missing data, data inconsistency etc. bring a lot of wrong information which will lead to wrong classify. It is impossible to build a convicive classifier with incompleteness or wrong data.

**Attribute quality.** Attributes provide information for classifying. The prediction accuracy can be improved by including more attributes. However, more attributes means calculating more attribute combinations and more training time. It is essential to choose attributes which are valuable for classification.

**Characteristics of the records to be predicted.** If Characteristics of the records to be predicted are different from records in training set, it may lead to high incorrect rate.

### 2.2. Missing Treatment Methods Using Classifier

In the popular methods for missing data handling, using classifier to predict and fill in missing data is a set of fast developing methods. Among the so many classifiers, which one is suitable and how to use is still in the research. Methods have been used are introduced as follow:

**K-Nearest Neighbor Imputation (KNN)** uses K-Nearest Neighbor algorithms to replace missing data. In the dataset, missing data may change the character of the record or relativeness with other records, then, probability of certain record be wrongly classified will rise.

**Internal Treatment Method for C4.5 (C4.5)** uses decision tree C4.5 to handle missing data. The internal treatment for missing data is one of C4.5’s advantages [6].

**Naïve Bayesian Imputation (NB)** uses naïve Bayesian classifier, one of the most widely used classifier, to predict and replace missing data.

Some researchers use Neural Network to predict missing data. But, building classifier on datasets containing missing data, no matter which one to use, the missing data will affect the process in certain extent, which depend on the specific classifier [5]. Missing data have little impact on certain classifier, that is to say, this classifier have higher prediction accuracy for datasets with missing data. The effect of missing data on classification algorithms can be used as a criterion to select suitable classifier.

### 3. Experiment and Analysis

#### 3.1. Sensitivity Analysis

Sensitivity analysis (SA) is to study the impacts of one or more input variables on the outputs of a model, that is, the sensitivity of the model to one parameter or a combination of parameters [8]. If a tiny change of an input leads to great changes of the output, the model is highly sensitive to that input. SA can help to identify the decisive input parameter of the model [8]. In our experiments, the proportion of missing data in the datasets is the parameter which affects the results of the classification models. The effect of missing data on the prediction accuracy will be investigated through the tiny changing of the missing rate.
3.2. Design of Experiments

This paper has selected 6 classifiers to study the influence of missing data to classifiers, that is, Naive Bayesian classifier (NB), Logistic Regression (LR), Neural Network which uses backpropagation to train (NN), K-Nearest Neighbours classifier (KNN), C4.5 decision-tree (C4.5) and classifier for building Logistic Model Trees (LMT). Ten datasets, which were collected from UCI, are used in the experiments, as shown in Table 1.

<table>
<thead>
<tr>
<th>No.</th>
<th>Datasets</th>
<th>Records</th>
<th>Attr.</th>
<th>Classes</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Breast</td>
<td>699</td>
<td>9</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>Bupa</td>
<td>345</td>
<td>6</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>Nursery</td>
<td>12960</td>
<td>8</td>
<td>5</td>
</tr>
<tr>
<td>4</td>
<td>German</td>
<td>1000</td>
<td>24</td>
<td>2</td>
</tr>
<tr>
<td>5</td>
<td>Crx</td>
<td>690</td>
<td>15</td>
<td>2</td>
</tr>
<tr>
<td>6</td>
<td>Pima</td>
<td>768</td>
<td>8</td>
<td>2</td>
</tr>
<tr>
<td>7</td>
<td>Vehicle</td>
<td>846</td>
<td>18</td>
<td>4</td>
</tr>
<tr>
<td>8</td>
<td>Cmc</td>
<td>1473</td>
<td>9</td>
<td>3</td>
</tr>
<tr>
<td>9</td>
<td>Ionosphere</td>
<td>351</td>
<td>34</td>
<td>2</td>
</tr>
<tr>
<td>10</td>
<td>Segment</td>
<td>2310</td>
<td>19</td>
<td>7</td>
</tr>
</tbody>
</table>

Table 1. Instruction of datasets

Three indexes are used to evaluate the missing influence on classifier: Prediction accuracy \( (Pa) \), Prediction profit \( (Pp) \) [10] and Prediction losing \( (Pl) \), which are defined as follows.

\[
Pa = \frac{\text{number of correctly predicted records}}{\text{number of all records}} \times 100\% \\
Pp = \frac{Pa - MD}{MD} \times 100\% \\
Pl = \frac{AC - Pa \text{ under certain missing rate}}{AC} \times 100\% 
\]

AC is the prediction accuracy without missing data. MD is the proportion of the class in the dataset which having the greatest number of records. Without any prediction model, if all the records are classified into that class, MD will be the prediction accuracy of the dataset. Prediction profit is proposed by Peng Liu, Elia El-Darzi et al. (2004) to evaluate the performance of the classifier [9].

Initially, each dataset is randomly divided into two parts: 2/3 records of the dataset are used as training subset, while the rest 1/3 as testing subset. Such process is repeated ten times and ten pairs of training subset and testing subset are generated. Then, a given percentage, 10%, 20%, …, 90% of missing data is artificially inserted into the training subsets at completely random. Finally, the six classification algorithms mentioned above are applied into the training subset to build up classifiers, and, these classifiers are used to classify instances in testing subset to investigate the classification accuracy. The average prediction accuracy of the six classifiers under different missing percentages is displayed from Figure 1 to Figure 6. Limited by the space, only part of the results is showed in this paper.

3.3. Results and analysis

See the figures, with the increase in missing rate, the prediction accuracies of all the classifiers have an obvious trend of decrease. In general, when the proportion of missing data in the dataset is less than 10%, they have little adverse impact on the classifiers. In the experiments, the average Prediction losing of six classifiers is 2.16% (see Table 2). If the missing rate is between 10% and 20%, the impact should not be neglected. The average Prediction losing rises to 4.63%. However, the adverse impact can be reduced significantly by some simple methods, such as replacing missing data by an approximation. If the missing rate exceeds 20%, there is an obvious decrease in the prediction accuracy and the missing data should be handled with high cautiousness. Appropriate methods should be chosen to eliminate the adverse impact of the missing data and optimize the performance of classifiers. In the real world, there are great quantities of missing data in databases and, usually, the proportion of missing data exceeds 20%.
However, if the proportion of missing data exceeds 50%, the average prediction losing rises to more than 10%. Obviously, the loss of prediction accuracy caused by missing data is quite huge. What’s more (Fig. 7), with the increase in the missing rate, the losing of prediction is rising with accelerated paces. That is to say, with the increase of quantity of the missing data, the little raising of the missing rate will result in a larger and larger decrease in the prediction accuracy.

The impact of missing data depends on classifiers. Among the six classifiers, Naive Bayesian classifier is the least sensitive to missing data, C4.5 is next to. And K-Nearest Neighbor susceptible to missing data most, Neural Network is next to. With the increasing in the missing rate, the prediction accuracy of Naive Bayesian classifier is almost the same as that with no missing data. Only when more than 70% data are missing, the prediction accuracy drops obviously. See Figure 7, the prediction losing of Naive classifier is always the lowest and that of the KNN is always the highest. Further more, the pace of KNN is also very fast. When missing rate exceeds 10%, there is an obvious and sharp increase in prediction losing. That means fewer missing data will have an obvious adverse impact on classifiers. Similar situation can be found in Figure 8. The prediction profit of Naive Bayesian classifier drops with the lowest and smoothest pace. And the K-Nearest Neighbor has the sharpest trend.

In summary, Naive Bayesian classifier is the least sensitive to missing data. Even if training subset has a lot of missing data, Naive Bayesian classifier can still make full use of the existing data and operate effectively. Among all the classifiers,
though the classification accuracy (with no data missing) of Naive Bayesian classifier isn’t the highest, it is the most adaptive to missing data. The Naive Bayesian classifier is recommended for the datasets with great quantities of missing data. While, selecting Naive Bayesian classifier to deal with missing data can also get a better solution.

4. Conclusion

Missing data may reduce the accuracy of prediction models. This paper mainly studies the impact of missing data to classification algorithms. The sensitivity of six representative classifiers to missing data is analyzed in the experiments. The results showed that, with the increasing of the missing rate, the classification accuracies of all the classification algorithms have an obvious trend of decrease. If the proportion of missing data exceeds 20%, there is an obvious decrease in the accuracy of prediction. Methods for missing data treatment should be chosen cautiously to eliminate the negative impact on the classification accuracy and optimize the performance of classifiers. Among the six classifiers, the Naive Bayesian classifier is the least sensitive to missing data and K-Nearest Neighbour is the most sensitive. In conclusion, for datasets suffered with missing data, we prefer to use Naive Bayesian classifier to deal with missing. Liu P. and Lei L et al. (2004) introduce missing data treatment methods based on Naive Bayesian classifier in detail [7]. And Naive Bayesian classifier can also provide a good accuracy without handling missing data.

Reference