A correlation based imputation method for incomplete traffic accident data

Rupam Deb1, Alan Wee-Chung Liew1, and Erwin Oh2

1School of Information and Communication Technology, Griffith University, Australia
rupam.deb@griffithuni.edu.au, a.liew@griffith.edu.au

2School of Engineering, Griffith University, Australia
y.oh@griffith.edu.au

Abstract. Death, injury and disability from road traffic crashes continue to be a major global public health problem. Recent data suggest that the number of fatalities from traffic crashes is in excess of 1.25 million people each year with non-fatal injuries affecting a further 20-50 million people. It is predicted that by 2030 road traffic accidents will have progressed to be the 5th leading cause of death and that the number of people who will die annually from traffic accidents will have doubled from current levels. Therefore, methods to reduce accident severity are of great interest to traffic agencies and the public at large. Road accident fatality rate depends on many factors and it is a very challenging task to investigate the dependencies between the attributes because of the many environmental and road accident factors. Any missing data in the database could obscure the discovery of important factors and lead to invalid conclusions. In order to make the traffic accident datasets useful for analysis, it should be preprocessed properly. In this paper, we present a novel method based on sampling of distributions obtained from correlation measures for the imputation of missing values to improve the quality of the traffic accident data. We evaluated our algorithm using two publicly available traffic accident databases of United States (explore.data.gov, data.opencolorado.org). Our results indicate that the proposed method performs significantly better than the three existing algorithms.

Keywords: Data mining; Data preprocessing; Decision tree; Data imputation; Traffic accident.

1 Introduction

A nation’s socio-economic development is highly dependent on the health status of its citizens. Road safety, which is mainly affected by road accident is said to be one of the major health concern. The burden of road accident causalities and damage is a major headache for both developed and developing countries. The global economic cost of road traffic accidents has been estimated at US$518b and has been calculated to account for 0.3% to 4% of the gross national product of many countries [12].

Digital data acquisition methods and storage technology have resulted in the growth of a huge amount of traffic data being stored in different types of databases. Large amount of traffic accident data has been collected with the advancement in sensor technologies [1]. Data sets often have missing values due to various reasons including equipment malfunctioning, human errors, and faulty data transmission. If an organization does not take extreme care during data collection, then approximately
5% or more missing/corrupt data may be introduced in the data sets [2, 6, 8-10]. Data preprocessing includes imputation of missing values, smoothing out noisy data, identification of incorrect data, and correction of inconsistent data. For missing value imputation, different imputation algorithms perform well on different types of datasets and missing values.

In this paper, we propose a simple but effective imputation algorithm called SCMI for imputing missing values on traffic accident datasets. Our propose algorithm measures the correlation after horizontally divide the large dataset depends on missing record(s). Our experiments show that the proposed algorithm has better imputation accuracy compared with several other existing algorithms. For validation, we compare our algorithm with some well-known missing value imputation techniques using three performance indicators: mean absolute error (MAE), coefficient of determination (CoD), and root mean square error (RMSE).

2 Background

Many missing value imputation approaches have been proposed recently for various applications [2-6, 11]. However, most of the existing methods are not suitable for both numerical and categorical missing values imputation [3].

For imputing numerical missing values, Expectation Maximization imputation (EMI) algorithm relies on estimating the mean and covariance matrices of the dataset. The EMI algorithm starts with initial estimates of the mean and the covariance matrix and cycles through the steps until the imputed values and the estimates of mean and covariance matrix stop changing appreciably from current iteration to the next iteration [4].

$k$-Nearest Neighbor based Imputation (kNNI) [5] method first finds user-defined $k$ number of records from the total dataset by using the Euclidean distance measure. kNNI is a simple method that performs well on the dataset having strong local correlation structure. In some existing techniques such as Local Weighted Linear Approximation Imputation (LWLA), Iterative Bi-Cluster based Local Least Square Imputation (IBLLS) methods, the $k$ value is estimated automatically [11]. However, the identification of a suitable value for $k$ can be a challenging task.

Rahman et al. proposed the Decision tree based imputation (DMI) [3] technique which uses the decision tree and EM algorithm for missing value imputation. They showed that DMI performed well compared with other existing imputation methods. However, for imputing categorical values, simple voting is used. Another more serious problem is that the authors did not define how the imputation is done if the missing values record falls in more than one leaves, a situation that could occur if there is more than one missing values in a record.

$k$-Decision tree based imputation (kDMI) [2] algorithm imputes missing values using two levels partitioning. Like DMI, kDMI algorithm also employed horizontal partitioning based on a decision tree in first level partitioning. For second level partitioning, the authors used a BestKNN approach to first find the best value of $k$ by searching all records of a leaf and calculated the root mean square error (RMSE) of
the non-missing attribute values. However, it is not clear if all the attributes of a record are categorical then how would RMSE being calculated using BestKNN. It is also not clear if the missing values record falls in more than one leaves, how the imputation would be done.

Here we propose SCMI, a correlation based missing value imputation technique to solve the above mentioned problems. We identify the relationships between the records using IS measure of the dataset and then impute missing values based on the strengths of correlation between records.

3 Proposed Approach

3.1 Basic Concept

We describe our algorithm using a toy dataset. First, the full dataset (DFull) is divided into two sub datasets. One subset contains records with missing values (DMiss) and other one without missing values (DComplete). A sample datasets DFull, DComplete, and DMiss are shown in TABLE 1, 2, and 3, respectively.

<table>
<thead>
<tr>
<th>Record</th>
<th>Driver status</th>
<th>Weather condition</th>
<th>Passenger number</th>
<th>Accident address</th>
<th>Injury severity</th>
</tr>
</thead>
<tbody>
<tr>
<td>R1</td>
<td>Drunk</td>
<td>Good</td>
<td>3</td>
<td>Sanders</td>
<td>Kill</td>
</tr>
<tr>
<td>R2</td>
<td>Drunk</td>
<td>Good</td>
<td>4</td>
<td>?</td>
<td>Kill</td>
</tr>
<tr>
<td>R3</td>
<td>Drunk</td>
<td>Good</td>
<td>2</td>
<td>Glendale</td>
<td>No injury</td>
</tr>
<tr>
<td>R4</td>
<td>Normal</td>
<td>Fair</td>
<td>3</td>
<td>Glendale</td>
<td>No injury</td>
</tr>
<tr>
<td>R5</td>
<td>Normal</td>
<td>Fair</td>
<td>?</td>
<td>Glendale</td>
<td>No injury</td>
</tr>
<tr>
<td>R6</td>
<td>?</td>
<td>Good</td>
<td>?</td>
<td>Glendale</td>
<td>Kill</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Record</th>
<th>Driver status</th>
<th>Weather condition</th>
<th>Passenger number</th>
<th>Accident address</th>
<th>Injury severity</th>
</tr>
</thead>
<tbody>
<tr>
<td>R1</td>
<td>Drunk</td>
<td>Good</td>
<td>3</td>
<td>Sanders</td>
<td>Kill</td>
</tr>
<tr>
<td>R3</td>
<td>Drunk</td>
<td>Good</td>
<td>2</td>
<td>Glendale</td>
<td>No injury</td>
</tr>
<tr>
<td>R4</td>
<td>Normal</td>
<td>Fair</td>
<td>3</td>
<td>Glendale</td>
<td>No injury</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Record</th>
<th>Driver status</th>
<th>Weather condition</th>
<th>Passenger number</th>
<th>Accident address</th>
<th>Injury severity</th>
</tr>
</thead>
<tbody>
<tr>
<td>R2</td>
<td>Drunk</td>
<td>Good</td>
<td>4</td>
<td>?</td>
<td>Kill</td>
</tr>
<tr>
<td>R5</td>
<td>Normal</td>
<td>Fair</td>
<td>?</td>
<td>Glendale</td>
<td>No injury</td>
</tr>
<tr>
<td>R6</td>
<td>?</td>
<td>Good</td>
<td>?</td>
<td>Glendale</td>
<td>Kill</td>
</tr>
</tbody>
</table>
Then, we match each record of \( D_{\text{Miss}} \) with \( D_{\text{Complete}} \) and create a table for each record of \( D_{\text{Miss}} \) using matched records of \( D_{\text{Complete}} \) and the corresponding missing record of \( D_{\text{Miss}} \). For example, R2 record, we get two records (R1, R3) matching of \( D_{\text{Complete}} \).

**Table 4. ASSIGN MISS DATASET FOR ACCIDENT ADDRESS**

<table>
<thead>
<tr>
<th>Record</th>
<th>Driver status</th>
<th>Weather condition</th>
<th>Passenger number</th>
<th>Accident address</th>
<th>Injury severity</th>
</tr>
</thead>
<tbody>
<tr>
<td>R1</td>
<td>Drunk</td>
<td>Good</td>
<td>3</td>
<td>Sanders</td>
<td>Kill</td>
</tr>
<tr>
<td>R3</td>
<td>Drunk</td>
<td>Good</td>
<td>2</td>
<td>Glendale</td>
<td>No injury</td>
</tr>
<tr>
<td>R2</td>
<td>Drunk</td>
<td>Good</td>
<td>4</td>
<td>?</td>
<td>Kill</td>
</tr>
</tbody>
</table>

Now, we select the possible impute value(s) from the non-missing records of each missing value assigned table. For example, in TABLE 4, to impute “Accident address” we have two possible imputed values ‘Sanders’ or ‘Glendale’. If there is more than one possible imputed value, the correlation between each possible imputed value and the matched non-missing value(s) is computed using IS measure. The IS measure between two items \( a \) and \( b \) is given by \( \text{IS}(a,b) = \frac{P(a,b)}{\sqrt{(P(a)P(b))}} \), where \( P \) denotes probability value [7]. For example, in TABLE 4, the IS measure of (‘Sanders’) and (‘Glendale’) have the same value of 1.

Here, from a total of four non-missing attributes values of R2 record, for ‘Sanders’ three attributes values are matched whereas for ‘Glendale’ two attributes values are matched and the ratios of matched attributes values for ‘Sanders’ and ‘Glendale’ are 0.75 and 0.5, respectively. The association between imputed value(s) and matched value(s) is stronger if IS measure value is bigger. Therefore, to give the priority of matched maximum number of attributes value(s), the ratio(s) and IS measure value(s) are multiplied to obtain a weighted IS. For example, for ‘Sanders’ the multiplied IS measure value is 0.75 and for ‘Glendale’ is 0.5.

Finally, the actual imputed value is obtained by random sampling from the list of possible imputed values based on the distribution of their weighted IS. Random sampling according to the weighted IS measure ensures that uncertainty in attribute values are modeled for and helps to reduce bias artefact in the imputed dataset.

### 3.2 Proposed Algorithm

**Declaration:**
- \( D_{\text{Full}} \): Full dataset, \( D_{\text{Complete}} \): Sub dataset without missing values, \( D_{\text{Miss}} \): Sub dataset with missing values, \( N \): Total number of non-missing records matched with a missing record in table \( T \), \( O_k \): possible imputed value(s) from the \( k \)-th record of \( T \), \( C_k \): \( k \)-th record weighted IS measure corresponding to \( O_k \)

**START**

**Step I:** Decompose full dataset into complete and missing values sub-datasets
- \( D_{\text{Full}} = D_{\text{Complete}} + D_{\text{Miss}} \)

**Step II:** Generate tables \( T \)
- FOR each missing record in \( D_{\text{Miss}} \) DO
Find records in D_complete that match with the non-missing attribute(s) value(s) of each missing record and generate T END FOR

**Step III:** Impute missing values in each T
FOR each T DO
FOR k = 1 to N
Q_k = number of matched attribute(s) / Total number of non-missing attribute(s) of the k-th matched record
O_k = possible imputed value(s) from the k-th matched record
C_k = (IS measure computed for O_k) * Q_k
END FOR
Imputed value(s) is obtained by randomly sampling from the set of possible imputed values \{ O_1…O_N \} based on the distribution specified by the set of IS measures \{ C_1…C_N \}
END FOR

**Step IV:** Return full dataset with missing values imputed
END

4 **Results and Discussion**

We compare the performance of our method with four imputation methods kDMI, DMI, kNNI, and EMI. We do experiment on 43 text file data (Large Truck Crash Causation Study File 1 and 2) and accident dataset of Denver County. The two Large Truck Crash Causation Study Files have different number of attributes and 92871 records. The Denver County has 89194 traffic accident records. The two Large Truck Crash Causation Study File datasets contain 3192 records with missing values. Denver County dataset contains 1902 records with missing values. Here, we first remove the records having missing values thereby create two complete datasets having 89679 and 87292 records without any missing values in two Large Truck Crash Causation Study Files and Denver County’s dataset.

We use four types of missing patterns [3]. In simple pattern, a record can have at most one missing value. In medium pattern, a record can have missing values for 2 – 50% of the total number of attributes. In a complex pattern, a record can have missing values for 51 – 80% of the total number of attributes. A blended pattern contains 25% records having missing values with simple pattern, 50% with medium pattern and 25% with complex pattern. We also use two types of missing models, namely overall and uniformly distributed (UD). In the UD missing model, each attribute has equal number of missing values. However, in the overall model, missing values are not equally distributed among the attributes and in the worst case all missing values can belong to a single attribute.

We artificially create missing values in the dataset by using 4 missing patterns, namely simple, medium, complex and blended, 4 missing ratios i.e. 2%, 4%, 8%, and 12% and two missing models, namely overall and uniformly distributed (UD). We have altogether 32 missing combinations (4 missing ratios, 4 missing patterns, 2 missing models). For each combination we generate 200 datasets.
We present the aggregate performances based on CoD for missing ratio, missing model and missing patterns on the traffic accident datasets in Figure 1. The figure shows that for all cases SCMI outperforms other methods in terms of CoD. In Figure 2, we present the aggregate performances based on RMSE for 4 missing ratios for two datasets. We present the performance of our algorithm based on CoD, RMSE and MAE for categorical missing value imputation in Figure 3.

Fig. 1. Average performance comparison based on coefficient of determination (CoD)

Fig. 2. Average performance for missing ratio comparison of SCMI and kDMI based on RMSE
Conclusion

In this paper, we proposed a new imputation method with the aim of analyzing traffic accidents data. Our algorithm imputes missing value using sampling based on distribution obtained from correlation measure, and has been shown to outperform several popular imputation methods on traffic accident data. In the future, we intend to run our technique on large and diverse traffic accident datasets.

References