Information Sciences xxx (2016) xxx–xxx

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

Missing value imputation for the analysis of incomplete traffic accident data

Q1 Rupam Deb*, Alan Wee-Chung Liew

School of Information and Communication Technology, Gold Coast Campus, Griffith University, QLD 4222, Australia

ARTICLE INFO

Article history:
Received 24 April 2015
Revised 30 December 2015
Accepted 2 January 2016
Available online xxx

Keywords:
Data preprocessing
Decision tree
Missing value imputation
Categorical data
Traffic accident

ABSTRACT

Death, injury and disability resulting 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. Both developed and developing countries suffer from the consequences of increase in human population, and therefore, vehicle population. Therefore, methods to reduce accident severity are of great interest to traffic agencies and the public at large. To analyse traffic accident factors effectively we need a complete traffic accident historical database. Any missing data in the database could prevent the discovery of important environmental and road accident factors and lead to invalid conclusions. In this paper, we present a novel imputation method that exploits the within-record and between-record correlations to impute missing data of numerical or categorical values. In addition, our algorithm accounts for uncertainty in real world data by sampling from a list of potential imputed values according to their affinity degree. We evaluated our algorithm using four publicly available traffic accident databases from the United States, the first of which is the largest open federal database (explore.data.gov) in the United States, and the second is based on the National Incident Based Reporting System (NIBRS) of the city and county of Denver (data.opencolorado.org). The other two are from New York's open data portal (Motor Vehicle Crashes-case information: 2011 and Motor Vehicle Crashes-individual information: 2011, data.ny.gov). We compare our algorithm with four state-of-the-art imputation methods using missing value imputation accuracy and RMSE. Our results indicate that the proposed method performs significantly better than the existing algorithms we compared.

1. Introduction

The high growth of the number of vehicles has led to roads with higher traffic density. The immediate effect of this situation is the dramatic increase of traffic accidents on the road, which has become a serious problem in many countries. For example, 2478 people died on Spanish roads in 2010, which means one death for every 18,551 inhabitants [7,8,11]. In the United States (according to the Department of Transportation, United States), in 2012, 33,561 people died in motor vehicle

* Corresponding author. Tel.: +61 426932694.
E-mail addresses: rupam.deb@griffithuni.edu.au, cse.rupam@gmail.com (R. Deb), a.liew@griffith.edu.au (A.W.-C. Liew).

http://dx.doi.org/10.1016/j.ins.2016.01.018
0020-0255/

traffic crashes. According to the Australian Bureau of Statistics, the majority of transport related deaths (72% in 2009) in Australia is associated with motor vehicles driven on public roads.

A nation's socio-economic development is highly dependent on the health status of its citizens. Road safety, which is mainly affected by road accidents is said to be one of the major health concerns. The burden of road accident causalities and damage is a major headache for both developed and developing countries. To reduce the number of road traffic accidents, it is necessary to characterize the causes of the accidents and also determine the factors which significantly affect the severity of injuries in road crashes. The global economic cost of road traffic accidents has been estimated at US$518b and has been calculated to account for 0.3–4% of the gross national product of many countries [15].

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 amounts of traffic accident data have been collected with the advancement in sensor technologies [34]. Using data mining technology, such as classification and clustering, we can uncover patterns of traffic activities and factors which lead to accident. The major reason that data mining has attracted a great deal of attention in information industry is due to the wide availability of huge amounts of data and the imminent need for turning such data into useful information and knowledge [28]. Data mining is the exploration and analysis of large dataset in order to discover knowledge and rules. Data mining is typically conceptualized as a three part process: preprocessing, learning, and post-processing.

Appropriate data preprocessing involves transforming the raw data into suitable format for subsequent analysis. To run the classification and clustering algorithms, there is a strong need for good data preprocessing to ensure the data are of good quality. Data preprocessing takes almost 80% of the total data mining effort [35]. It is also known that good results can be achieved by using data mining algorithms only if there is a good quality dataset [19].

Real-life data are frequently imperfect: erroneous, incomplete, uncertain and vague. Datasets 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 into the datasets [10,20,33]. Data preprocessing includes imputation of missing values, smoothing out noisy data, identification of incorrect data, and correction of inconsistent data. In this work, we propose a novel imputation algorithm to address the problem of missing attribute values.

Specifically, we propose a new decision tree and sampling based missing value imputation algorithm called DSMI for missing value imputation. Our algorithm measures the correlation after horizontally dividing the large dataset depending on the missing record(s). We divide the whole dataset on different horizontal segments based on non-missing attributes of the missing records and impute the missing values based on sampling from the distributions induced by the correlation. 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 imputation accuracy and RMSE.

In our study, we use two truck crash datasets, the accident dataset of Denver County, and two crash datasets from New York's open data portal. Most of the columns (attributes) of these records are categorical. The domain of the categorical data is represented by nominal, ordinal and interval based variables. Nominal variables have values that have no natural ordering (e.g. airbag conditions: Ruptured, Cut, Torn); Ordinal variables do have a natural order (e.g. day of the week); and Interval variables are created from intervals on a contiguous scale (e.g. age group 13–19).

This paper is organized as follows: in Section 2 we present a literature review of related work. Our proposed technique is described in Section 3. Experimental results are discussed in Section 4. Finally Section 5 draws the concluding remarks.

2. Related work

Imputation of missing values is an important data mining task for improving the quality of the data mining result. Many missing value imputation algorithms have been proposed for various applications [1,3,4,14,16–18,23–27,30]. Some of these methods are: expectation maximization imputation (EMI) [27], decision tree based missing value imputation (DMI) [26], similarity based missing value imputation (SiMI) [26], *k*-decision tree based missing value imputation (*k*DMI) [25], *k*-nearest neighbour based imputation (kNNI) [1], local weighted linear approximation imputation (LWLA) [18], and framework for imputing missing values using co-appearance, correlation and similarity analysis (FIMUS) [24].

To impute numerical missing values, the EMI algorithm [27] relies on estimating the mean and covariance matrix of the dataset. The EMI algorithm starts with an initial estimate of the mean and the covariance matrix, and iterates until the imputed values and the estimates of mean and covariance matrix stop changing appreciably from the current iteration to the next iteration [9,27]. The EMI algorithm is only applicable to datasets in which the missing values are missing at random. The main drawback of this method is that for imputing the missing value, EMI algorithm uses information from the whole dataset and therefore is suitable only for datasets that exhibit strong correlations between attributes.

Instead of using information from the whole dataset, kNNI method [1] imputes missing values using *k* number of similar records. This method first finds user-defined *k* number of records from the total dataset by using the Euclidean distance measure. To impute a numerical missing value, the method utilizes the mean value of the specific attribute within the *k* most similar records of the entire dataset. If the missing attribute is categorical, then the method utilizes the most frequent value of the attribute within the *k* most similar records. kNNI is a simple method that performs well on dataset that has strong local correlation structure. However, the method can be expensive for large dataset since for each record with missing

Please cite this article as: R. Deb, A.W.-C. Liew, Missing value imputation for the analysis of incomplete traffic accident data, Information Sciences (2016), http://dx.doi.org/10.1016/j.ins.2016.01.018