DATA CLEANING FOR DATA ANALYSIS

Huang Shan, E. Gubin
Tomsk Polytechnic University
gubine@tpu.ru

Introduction

Data preprocessing step is often ignored. However, it is a crucial step in data mining process, because if you put garbage in, you get garbage out. Data gathering is not always strictly controlled, so the data usually contains such imperfections as missing values, odd values (Age: -10), impossible data combinations (Gender: Male, Pregnant: Yes) etc. Performing analysis on the data that has not been preprocessed may lead to problems and misleading results. That is why the quality of the data is first.

If there is much irrelevant and redundant information or unreliable data, the quality of consequent analysis is nothing but poor. Thus, although this step may take considerable amount of time, it should not be omitted. Data preprocessing includes cleaning, normalization, transformation, feature extraction and selection, etc. The object of study is credit scoring dataset. The subject of study are methods of data cleaning.

The goal of the study is to develop and investigate algorithm of data cleaning for data analysis.

1. Formulation of the problem

Raw data is highly susceptible to noise, missing values, and inconsistency. The quality of data affects the data mining results. In order to help improve the quality of the data and, consequently, of the mining results raw data is preprocessed to improve the efficiency and ease of the mining process. Data preprocessing is one of the most critical steps in a data mining process, which deals with the preparation and transformation of the initial dataset. Data preprocessing methods are divided into following categories [1]: 1)data cleaning, 2)data integration, 3)data transformation and 4)data reduction.

Data cleaning routines work to clean the data by filling in missing values, smoothing noisy data, identifying or removing outliers, and resolving inconsistencies. Although most mining routines have some procedures for dealing with incomplete or noisy data, they are not always robust. Instead, they may concentrate on avoiding overfitting the data to the function being modeled. The tasks of data cleaning are: 1)missing values imputation, 2) identifying outliers, 3) correction inconsistent data, 4) resolving redundancy caused by data integration

Incomplete data can occur for a number of reasons. Attributes of interest may not always be available, such as customer information. Other data may not be included simply because it was not considered important at the time of entry. Relevant data may not be recorded due to a misunderstanding, or because of equipment malfunctions. Data that were inconsistent with other recorded data may have been deleted. Furthermore, the recording of the history or modifications to the data may have been overlooked. Missing data, particularly for tuples with missing values for some attributes, may need to be inferred. Data can be noisy, having incorrect attribute values, owing to the following. The data collection instruments used may be faulty. There may have been human or computer errors occurring at data entry. Errors in data transmission can also occur. There may be technology limitations, such as limited buffer size for coordinating synchronized data transfer and consumption. Incorrect data may also result from inconsistencies in naming conventions or data codes used. Duplicate tuples also require data cleaning.

Missing Data

Depending on what causes missing data, the gaps will have a certain distribution. Understanding this distribution may be helpful in two ways. First, it may be employed as background knowledge for selecting an appropriate imputation algorithm. Second, this knowledge may help to design a reasonable simulator, that removes missing data from a test set. Such a simulator will help to generate data where the true values (i.e., the potentially ideal imputation data) is known. Hence, the quality of an imputation algorithm can be tested [2].

Missing data mechanisms can be divided into three categories: 1) missing completely at random (MCAR), 2) missing at random (MAR), 3) Missing not at random (MNAR)

In practice, assigning data gaps to a category can be blurry, because the underlying mechanisms are simply unknown. While MAR and MNAR diagnosis needs manual analysis of the patterns in the data and application of domain knowledge, MCAR can be tested for with t-test or Little s test [3]. The vast majority of missing data methods require MAR or MCAR, since the missing data mechanism is said to be ignorable for them [4]. Since MAR enables imputation algorithms to employ correlations with other variables, algorithms can achieve better results than for MCAR. MNAR is called non-ignorable, because in order to do the imputation a special model for why data is missing and what the likely values are has to be included.

Outliers

Outliers are extreme values that deviate from other observations on data, they may indicate a variability in a measurement, experimental errors or a novelty. In other words, an outlier is an observation that diverges from an overall pattern on a sample.

Outliers can be of two kinds: 1) univariate,

2) multivariate

Univariate outliers can be found by looking at a distribution of values in a single feature space. Multivariate outliers can be found in a n-dimensional space (of n-features).

Outliers can also come in different flavors, depending on the environment: 1) point outliers, 2) contextual outliers, 3) collective outliers.

Point outliers are single data points that lay far from the rest of the distribution. Contextual outliers can be noise in data, such as punctuation symbols when realizing text analysis or background noise signal when doing speech recognition. Collective outliers can be subsets of novelties in data such as a signal that may indicate the discovery of new phenomena.

2. Overview of current solutions

Missing data, discarding data

Many missing data approaches simplify the problem by throwing away data. These approaches may lead to biased estimates. In addition, throwing away data can lead to estimates with larger standard errors due to reduced sample size.

Complete-case analysis

A direct approach to missing data is to exclude them. In the regression context, this usually means complete-case analysis: excluding all units for which the outcome or any of the inputs are missing.

Two problems arise with complete-case analysis:

1) If the units with missing values differ systematically from the completely observed cases, this could bias the complete-case analysis. 2) If many variables are included in a model, there may be very few complete cases, so that most of the data would be discarded for the sake of a simple analysis.

Available-case analysis

Another simple approach is available-case analysis, where different aspects of a problem are studied with different subsets of the data.

This approach has the problem that different analyses will be based on different subsets of the data and thus will not necessarily be consistent with each other. In addition, as with complete-case analysis, if the nonrespondents differ systematically from the respondents, this will bias the available-case summaries. Available-case analysis also arises when a researcher simply excludes a variable or set of variables from the analysis because of their missing-data. In a causal inference context (as with many prediction contexts), this may lead to omission of a variable that is necessary to satisfy the assumptions necessary for desired (causal) interpretations.

Univariate imputation

Rather than removing variables or observations with missing data, another approach is to fill in or "impute" missing values. A variety of imputation approaches can be used that range from extremely simple to rather complex. These methods keep the full sample size, which can be advantageous for bias and precision; however, they can yield different kinds of bias, as detailed in this section. Whenever a single imputation strategy is used, the standard errors of estimates tend to be too low. The intuition here is that we have substantial uncertainty about the missing values, but by choosing a single imputation we in essence pretend that we know the true value with certainty.

Mean imputation

Perhaps the easiest way to impute is to replace each missing value with the mean of the observed values for that variable. Unfortunately, this strategy can severely distort the distribution for this variable, leading to complications with summary measures including, notably, underestimates of the standard deviation. Moreover, mean imputation distorts relationships between variables by "pulling" estimates of the correlation toward zero.

Last observation carried forward

In evaluations of interventions where pre-treatment measures of the outcome variable are also recorded, a strategy that is sometimes used is to replace missing outcome values with the pre-treatment measure. This is often thought to be a conservative approach (that is, one that would lead to underestimates of the true treatment effect). However, there are situations in which this strategy can be anticonservative. For instance, consider a randomized evaluation of an intervention that targets couples at high risk of HIV infection. From the regression-to-the-mean phenomenon, we might expect a reduction in risky behavior even in the absence of the randomized experiment; therefore, carrying the last value forward will result in values that look worse than they truly are. Differential rates of missing data across the treatment and control groups will result in biased treatment effect estimates that are anticonservative.

3. Conclusion

The proposed algorithm includes applying different methods to the given data and comparing their performance by further accuracy check of classification algorithms applied to the cleaned data.

Bibliography

1. Anshu B. Data Preprocessing Techniques for Data Mining // Data Mining Techniques and Tools for Knowledge Discovery in Agricultural Datasets. New Delhi, 2011. P. 6.
2. Moritz S. et al. Comparison of different Methods for Univariate Time Series Imputation in R // CoRR. 2015. Vol. abs/1510.03924.
3. Little R.J.A. A Test of Missing Completely at Random for Multivariate Data with Missing Values // J. Am. Stat. Assoc. Taylor & Francis, 1988. Vol. 83, № 404. P. 1198–1202.
4. RUBIN D.B. Inference and missing data // Biometrika. 1976. Vol. 63, № 3. P. 581–592.