International Journal For Technological Research In Engineering
Vol. 1, Issue. 2, Oct – 2013 *ISSN (Online): 2347 - 4718*

A HEURISTIC APPROACH FOR CLEANING DISGUISED MISSING DATA

Prof. Bhargav Modi[1], Prof. Satvik Khara[2]
[1]*Department of Computer Engineering. Silver Oak College of Engineering & Technology, Ahmedabad*
[2]*Department of IT, Silver Oak College of Engineering & Technology, Ahmedabad*

Abstract: In some applications such as filling in a customer information form on the web, some missing values may not be explicitly represented as such, but instead appear as potentially valid data values. Such missing values are known as dis- guised missing data, which may impair the quality of data analysis severely, such as causing significant biases and misleading results in hypothesis tests, correlation analysis and regressions. The very limited previous studies on cleaning disguised missing data use outlier mining and distribution anomaly detection. They highly rely on domain background knowledge in specific applications and may not work well for the cases where the disguise values are inliers. To tackle the problem of cleaning disguised missing data, in this paper, we first model the distribution of disguised missing data, and propose the embedded unbiased sample heuristic. Then, we develop an effective and efficient method to identify the frequently used disguise values which capture the major body of the disguised missing data. Our method does not require any domain background knowledge to find the suspicious disguise values. We report an empirical evaluation using real data sets, which shows that our method is effective the frequently used disguise values found by our method match the values identified by the domain experts nicely. Our method is also efficient and scalable for processing large data sets.

Keywords- Data Quality, Data Cleaning, Disguised Missing Data

I. INTRODUCTION

Anyone who does statistical data analysis or data cleaning of any kind runs into the problems of missing data. In a characteristic dataset we always land up in some missing values for attributes. For example in surveys people generally tend to leave the field of income blank or sometimes people have no information available and cannot answer the question. Also in the process of collecting data from multiple sources some data may be inadvertently lost. For all these and many other reasons, missing data is a universal problem in both social and health sciences. This is because every standard statistical method works on the fact that every problem has information on all the variables an it needs to be analyzed. The most common and simple solution to this Data Cleaning is if any case has missing data for any of the attribute to be analyzed we can simply ignore it. This will give us a dataset which will not contain any missing value and we can then use any standard methods to process it further. But this method has a major drawback which is deleting missing values sometimes might lead to ignoring a large section of the original sample. This paper first illustrates different types of missing values and analyzes their consequences on datasets. After that we study two approaches taken by researchers to identify missing data from datasets in different scenarios.

2. DIFFERENT TYPE MISSING DATA

The problem of missing data resides in almost all the surveys and designed experiments. As stated before one of the common method is to ignore cases of missing values. Ignoring cases of missing values may sometimes lead to elimination of a major portion of the dataset thus leading
Into inappropriate results. The different types of missing mechanisms are stated as below:

With some sorts of research it is not unusual to have cases for which there are missing data for some but not all variables. There may or there may not be a pattern to the missing data. The missing data may be classified as MCAR, MAR, or MNAR.

Missing Not at Random (MNAR)

Some cases are missing scores on our variable of interest. Y
Suppose that Y is the salary of faculty members.
Missingness on Y is related to the actual value of Y.
Of course, we do not know that, since we do not know the values of Y for cases with missing data.
For example, faculty with higher salaries may be more reluctant to provide their income.
If we estimate mean faculty salary with the data we do have on hand it will be a biased estimate.
There is some mechanism which is causing missingness, but we do not know what it is.

Missing At Random (MAR)

Missingness on Y is not related to the true value of Y itself or is related to Y only through its relationship with another variable or set of variables, and we have scores on that other variable or variables for all cases

For example, suppose that the higher a professor's academic rank the less likely e is to provide e's salary. Faculty with higher ranks has higher salaries. We know the academic rank of each respondent.

We shall assume that within each rank whether Y is missing or not is random – of course, this may not be true, that is, within each rank the missingness of Y may be related to the true value of Y.
Again, if we use these data to estimate mean faculty salary, the estimate will be biased. However, within conditional distributions the estimates will be unbiased – that is, we can estimate without bias the mean salary of lecturers, assistant professors, associate professors, and full professors. We might get an unbiased estimate of the overall mean by calculating a weighted mean of the conditional means where GM is the estimated grand mean, pi is, for each rank, the proportion of faculty at that rank, and Mi is the estimated mean for each rank.

Missing Completely at Random (MCAR)

There is no variable, observed or not, that is related to the amusingness of Y. This is probably never absolutely true, but we can pretend that it is.

3. FINDING PATTERN OF MISSING DATA

The MVA module mentioned in Tabachnik and Fidell is an add-on that is sometimes included with SPSS, sometimes not. Curiously, the version of SPSS (20) distributed to faculty to use on campus at ECU does not contain the MVA and multiple imputation modules, but that distributed for off campus use does. You can, however, obtain the same t tests that are produced by MVA without having that module. Suppose that the variable of interest is income and you are concerned only with how missingness on income is related with scores on the other variables. Create a missingness dummy variable (0 = not missing the income score, 1 = missing the income score). Now use t tests (or Pearson r) to see if missingness on income is related to the other variables (MAR).

4. DEALIING WITH THE PROBLEM OF MISSING DATA

Deletion of Cases.

Delete from the data analyzed any case that is missing data on any of the variables used in the analysis. If there are not a lot of cases with missing data and you are convinced that the missing data are MCAR, then this is an easy and adequate solution. Estimates will not be biased. Of course, the reduction of sample size results in a loss of power and increased error in estimation (wider confidence intervals).

Deletion of Variables.

Delete any variable that is missing data on many cases. This is most helpful when you have another variable that is well related to the troublesome variable and on which there are not many missing values.

Mean Substitution

For each missing data point, impute the mean on that variable. "Imputation" is the substitution of an estimated value in this case a mean) for the missing value. If the cases are divided into groups, substitute the mean of the group in which the case belongs. While this does not alter the overall or group means, it does reduce the overall or group standard deviations, which is undesirable.

Missingness Dummy Variable

Maybe the respondents are telling you something important by not answering one of the questions. Set up a dummy variable with value 0 for those who answered the question and value 1 for those who did not. Use this dummy variable as one of the predictors of the outcome variable. You may also try using both a missingness dummy variable and the original variable with means imputed for missing values.

Regression

Develop a multiple regression to predict values on the variable which has missing cases from the other variables. Use the resulting regression line to predict the missing values. Regression towards the mean will reduce the variability of the data, especially if the predicted variable is not well predicted from the other variables.

Multiple Imputations

A random sample (with replacement) from the data set is used to develop the model for predicting missing values. Those predictions are imputed. A second random sample is used to develop a second model and its predictions are imputed. This may be done three or more times. Now you have three or more data sets which differ with respect to the imputed values. You conduct your analysis on each of these data sets and then average the results across data sets

Pairwise Correlation Matrix

For the variables of interest, compute a correlation matrix that uses all available cases for each correlation. With missing data it will not be true that all of the correlations will be computed on the same set of cases, and some of the correlations may be based on many more cases than are others. After obtaining this correlation matrix, use it as input to the procedure that conducts your analysis of choice. This procedure can produce very strange results.

5. MISSING ITEM DATA WITHIN A UNIDIMENSIONAL SCALE

Suppose that you have a twenty item scale that you trust is

uni-dimensional. You have done item analysis and factor analysis that supports your contention that each item (some of them first having been reflected) measures the same construct (latent variable). Some of your subjects have failed to answer one or more of the items on your scale.

What to do?

A relatively simple solution is to compute, for each subject, the mean of that subject's scores on the items he or she did answer on that scale. If all of the items are measuring the same construct (with the same metric), that should give you a decent estimate of the subject's standing on that construct. Please note that this is not what is commonly referred to as "mean substitution." With "mean substitution," one substitutes for the missing item score the mean score from other subjects on that item

Letting "miss" be the variable that is number of items with missing data on the scale: data cull; set Alicia; if Miss < 3;

If you are working with multiple scales, it might be easier to use an If, then, statement to set to missing the scale scores of any subject with too many missing item scores:

If Miss > 2, then misanth =. ; Else misanth = MEAN (of q1-q20);

If you are using SPSS, use the MEAN.N function. For example Compute misanth = mean.18 (Q1 to Q20). The number after "mean." specifies how many item scores must be nonmissing for the scale score to be computed.

6. ANALYSIS

Normality

If you have a large sample to tests may be significant when the deviation from normality is trivial and of no concern, and if you have a small sample to tests may not be significant when you have a big problem. If you have a problem you may need to transform the variable or change the sort of analysis you are going to conduct.

Homogeneity of Variance

Could these sample data have come from populations where the group variances were all identical? If the ratio of the largest group variance to the smallest group variance is large, be concerned. There is much difference of opinion regarding how high that ratio must be before you have a problem. If you have a problem you may need to transform the variable or change the sort of analysis you are going to conduct.

Homoscedasticity

This homogeneity of variance assumption in correlation/regression analysis. Careful inspection of the residuals should reveal any problem with heteroscedasticity. Inspection of the residuals can also reveal problems with the normality assumption and with curvilinear effects that you had not anticipated. See Bivariate Linear Regression and Residual Plots. If you have a problem you may need to transform the variable or change the sort of analysis you are going to conduct.

Sphericity

This assumption is made with univariate-approach correlated samples ANOVA. Suppose that we computed difference scores (like those we used in the correlated t test) for Level 1 vs. Level 2, Level 1 vs. Level 3, Level 1 vs. Level 4, and every other possible pair of levels of the repeated factor. The sphericity assumption is that the standard deviation of each of these sets of difference scores (1 vs. 2, 1 vs. 3, etc.) is a constant in the population. Mauchley's test is used to evaluate this assumption. If it is violated there you can stick with the univariate approach but reduce the df, or you can switch to the multivariate-approach which does not require this assumption.

7. CONCLUSION

Data cleaning and preparation is the primary step in data mining process. We first identify different types of missing data and then discuss two approaches to deal with missing data in different scenarios. This paper addresses the issues of handling missing values in datasets and methods in which missing values can be tackled. We first discuss the different types of missing Data and analyze their impact on the dataset

REFERENCES:

[1] Judi Scheffer, 2002. Dealing with Missing Data, Res. Lett.Inf. Math.Sci (2002). Quad A, Massey University, P.O. Box 102904 N.S.M.C, Auckland, 1310.

[2] Popova, V. 2006. Missing Values in Monotone Data Sets. In Proceedings of the Sixth international Conference on intelligent Systems Design and Applications (Isda'06) - Volume 01 (October 16 - 18, 2006). ISDA. IEEE Computer Society, Washington, DC, 627-632. DOI= http://dx.doi.org/10.1109/ISDA.2006.195

[3] Pearson, R. K. 2006. The problem of disguised missing data. SIGKDD Explor. Newsl. 8, 1 (Jun. 2006), 83-92. DOI= http://doi.acm.org/10.1145/1147234.1147247

[4] Hua, M. and Pei, J. 2007. Cleaning disguised missing data: a heuristic approach. In Proceedings of the 13th ACM SIGKDD international Conference on Knowledge Discovery and Data Mining (San Jose,California, USA, August 12 - 15, 2007). KDD '07. ACM, New York, NY, 950-958. DOI= http://doi.acm.org/10.1145/1281192.1281294

[5] Calders, T., Goethals, B., and Mampaey, M. 2007.
Mining itemsets in the presence of missing values.
In Proceedings of the 2007 ACM Symposium on Applied
Computing (Seoul, Korea, March 11 - 15, 2007).
SAC '07. ACM, New York, NY, 404-408. DOI=
http://doi.acm.org/10.1145/1244002.1244097