The impact of cleanliness of data on Machine Learning analysis

Shubhangi Kaushik

Abstract

The goal of the project:- Studying the influence of the cleanliness of a data set(MTBLS90) on the working and predictability of the ML methods. The corruption in the data using missing and incorrect values were set at 10 and 50 percentage.

Main results of the project:- The corrupted data sets were not compatible with the algorithms used due to the presence of missing and incorrect value. Cleaning the data using imputation shows improvement in the working of ML methods.

The project helped in understanding how to spot incorrect data within the data set, along with null values. It helped in exploring the different methods of imputing/rectifying dirty values, to clean the data set of incorrect/null values. The Analysis gave a glimpse of how imperfections in data left unresolved, impacts the performance and the results of any good and efficient ML methods.

Goal

The goal of the project was to conduct data analysis, extract important metabolites and compare results using two ML methods for the different versions of a data set(MTBLS90).The original version of data contains no incorrect or missing values, however, 4 versions constructed from the original, contain 10 and 50 percentage of missing/incorrect values. The dirty data versions were cleaned, analyzed, and compared with the original data set.

Data

For the analysis, the metabolomics data for the project ID: MTBLS90 was collected from Metabolights database, composed of the serum samples for people above age 70.[3].

The data set consists of 2 sheets, the first sheet named “Data Sheet” has information on 189 metabolite concentrations and their associated columns like “Idx”, “SampleID”, “Class” representing the sex of the sample, and “Sex”. The second sheet named “Peak Sheet” contains the metadata related to each metabolite from the Data Sheet.

Steps for generating corrupt data using NaN for 10 and 50 percent of data

The process of replacing the accurate values of the initial dataset with NaN involved generating random indices for both rows and columns. Subsequently, the genuine values for those indices were replaced with NaN at these positions.[2][1] Eventually, 2 corrupted data sets were created for NaN, where the first data set contained 10 percent of NaN and the other 50 percent.

Steps for generating corrupt data using incorrect data for 10 and 50 percent of data

Random values were first generated within the normalized range of the dataset. They were then either added or subtracted to the actual values found at the randomly generated indexes for the rows and columns. This random index generation was done in the same way as for the NaN corrupted data generation. Eventually, 2 more data sets were obtained, where one contained 10 percent of incorrect data, and the second contained 50 percent of incorrect data.

Data Preprocessing

Steps for generating clean data

For NaN corrupted data set

For generating clean data sets, the data sets containing NaN values at strength 10 and 50 percent were imputed using mean and median. To impute class values, the plan was to impute the values using logistic regression, however, the correlation of the class with the metabolites turned out to be very less. Also, imputing class values is not valid through mean or median because of it being categorical data with binary values. Thus, instead of imputing the rows with the class values as NaN, they were dropped.

For incorrect data corrupting data set

For imputing the incorrect values, first, the upper and lower limit of the data were calculated to determine the outliers followed by imputing them through the median. The imputation of the values using mean is not viable here, as the mean value obtained were getting highly influenced by the outliers Ultimately, a total of 8 clean data sets were obtained.

Data Analysis

From 4 corrupted data sets, 8 clean data sets using Imputation methods like Mean imputation and median imputation were obtained.

Mean Imputation:- Here the imputation of the missing values is done by taking the mean of the non missing values. Moreover, this method is very sensitive to outliers. Median Imputation:-the Imputation of the missing values is done by taking the median of the non missing data. This method is very robust, as it is not swayed by the magnitude of outliers.

All these 12 data sets were compared with the 1 original data set using Support Vector Machine(SVM) linear and Principal Component Regression(PCR) methods.

These ML methods use the metabolite data to predict the class labels:-
Support Vector Machine(SVM)linear:- This method finds the best hyperplane to divide two groups by adjusting the margin in a way that the distance is minimum for all the groups from the hyperplane. The value for the hyperparameter C defines this hyperplane.
Principal Component Regression(PCR):- Principal Component Analysis(PCA) identifies the most important directions of variation in the data, which are represented by the principal components. The gaps between groups along these components can provide insights into the dissimilarity or separation between those groups. Regression is performed by the algorithm between the components and the dependent variable.

After the preprocessing step, the data sets were divided into test and train sets with 33.33% of the data in the test and the rest in the train sets. The class column is the labels, whereas the metabolite data are the predictor variables[4].
The top ten metabolites were obtained for the ML methods mentioned above. The results for all the data sets were compared and the analysis was conducted on how the corruption of the data impacts the accuracy of the ML methods in predicting the metabolites. Additionally, the study investigated how cleaning the corrupted data could lead to improved results.

Results

After running SVM linear and PCR on the 13 data sets(1 original, 4 corrupted, 8 cleaned data sets), the graphs were obtained.

The original data set worked fine on the ML methods used. However, when the NaN corrupted data sets were used, an error stating NAN values found was generated at knnImpute. SVM and PCR threw the same error when KnnImpute was removed.

When ’incorrect data’ was passed in the mentioned algorithms, an error was generated stating NaN values were found, as the logarithmic conversion of negative values generates NaN values.

The algorithms worked fine when applied to the cleaned NaN data set with mean impute and median impute.
On trying to clean the ’incorrect data’ using median imputation, it took a long time. It was however able to run for the subset of the data.

Discussion

The corrupted data sets used in the mentioned ML methods severely affected the performance. It was mainly because the NaN values could not be processed in these methods. The NaN in the ’incorrect data’ came from the logarithmic conversion of the negative values.

Imputing the dataset corrupted by NaN values was found to be easier compared to the dataset corrupted by incorrect values. This is due to the fact that incorrect values are typically more challenging to detect. In the conducted analysis, although the incorrect values were identified, the imputation process took a considerable amount of time due to computational limitations on our systems.

Influence of different types of data on Machine Learning analysis		
TYPE OF DATA	STATUS	REASON FOR ”Fail”
Original data (MTBLS90)	Success	-
10% null values introduced to MTBLS90	Fail	knnimpute function, and PCR and SVM algorithms cannot work with the presence of NaN.
10% null values introduced to MTBLS90 replaced by Mean	Success	-
10% null values introduced to MTBLS90 replaced by Median	Success	-
50% null values introduced to MTBLS90	Fail	knnimpute function, and PCR and SVM algorithms cannot work with the presence of NaN.
50% null values introduced to MTBLS90 replaced by Mean	Success	-
50% null values introduced to MTBLS90 replaced by Median	Success	-
10% noise values introduced to MTBLS90	Fail	Log conversion was converting the negative values to NaN.
10% noise values introduced to MTBLS90, outliers replaced by Median	No Status	High time complexity,
50% noise values introduced to MTBLS90	Fail	Log conversion was converting the negative values to NaN.
50% noise values introduced to MTBLS90, outliers replaced by Median	No Status	High time complexity

Table 1 The table outlines different data sets that were analyzed and the issues encountered.

The metabolites obtained from the original data set were compared to the metabolites obtained from the mean and median imputed NaN value data sets as given in Table 2. For the data set with 10 percent NaN values, the metabolites obtained were similar to the original dataset. However, when compared to the data set with 50 percent NaN values, some metabolites were different. The 50 percent of data corruption has a bigger impact on the algorithms’ working and predicted metabolites.

Important Metabolites found from different data sets						
Id	Label	Original Dataset	10% null (mean)	10% null (median)	50% null (mean)	50% null (median)
M10	Creatine	Yes	Yes	Yes	Yes	No
M2	Creatinine	Yes	Yes	Yes	Yes	Yes
M141	Ceramide phosphoethanolamine(35:2) Sphingomyelin	Yes	Yes	Yes	Yes	Yes
M142	Phosphatidylcholine	Yes	Yes	Yes	Yes	Yes
M152	sphingomyelin	Yes	Yes	Yes	No	Yes
M15	2-ketohexanoic acid	Yes	Yes	Yes	No	No

Table 2 It was examined whether the common metabolites with the top 10 coefficient scores detected by SVM and PCR in the original data set were also present in the other data sets.

Appendix

Author details

References

1. Data cleaning.
2. Ml data cleaning.
3. Erik Ingelsson Lars Lind Andrea Ganna, Samira Salihovic. Mtbls90: Large-scale non-targeted serum metabolomics in the prospective investigation of the vasculature in uppsala seniors. *EBI*.
4. Stacey N. Reinke David I. Broadhurst Kevin M. Mendez. A comparative evaluation of the generalised predictive ability of eight machine learning algorithms across ten clinical metabolomics data sets for binary classification. *Metabolics*, 2019.