Data Modelling and Pipelining

MA19M020

November 2020

1 Aim:

Medical datasets are usually affected by several problems, such as missing values, inconsistencies, redundancies, that can influence the data mining process and the extraction of useful knowledge. Poor quality of the collected clinical data, in terms of incomplete, incorrect or improper values, can produce detrimental consequences: as an example, incorrect calculation of outcome prediction might lead to an improper medical treatment for the patient For these reasons, a preprocessing phase should be performed for improving the overall quality of data and, consequently, of the information that may be discovered from them.Below i have discussed some steps of data preprocessing to improve the quality of a large dataset derived from clinical trials.

2 Data Pipeline:

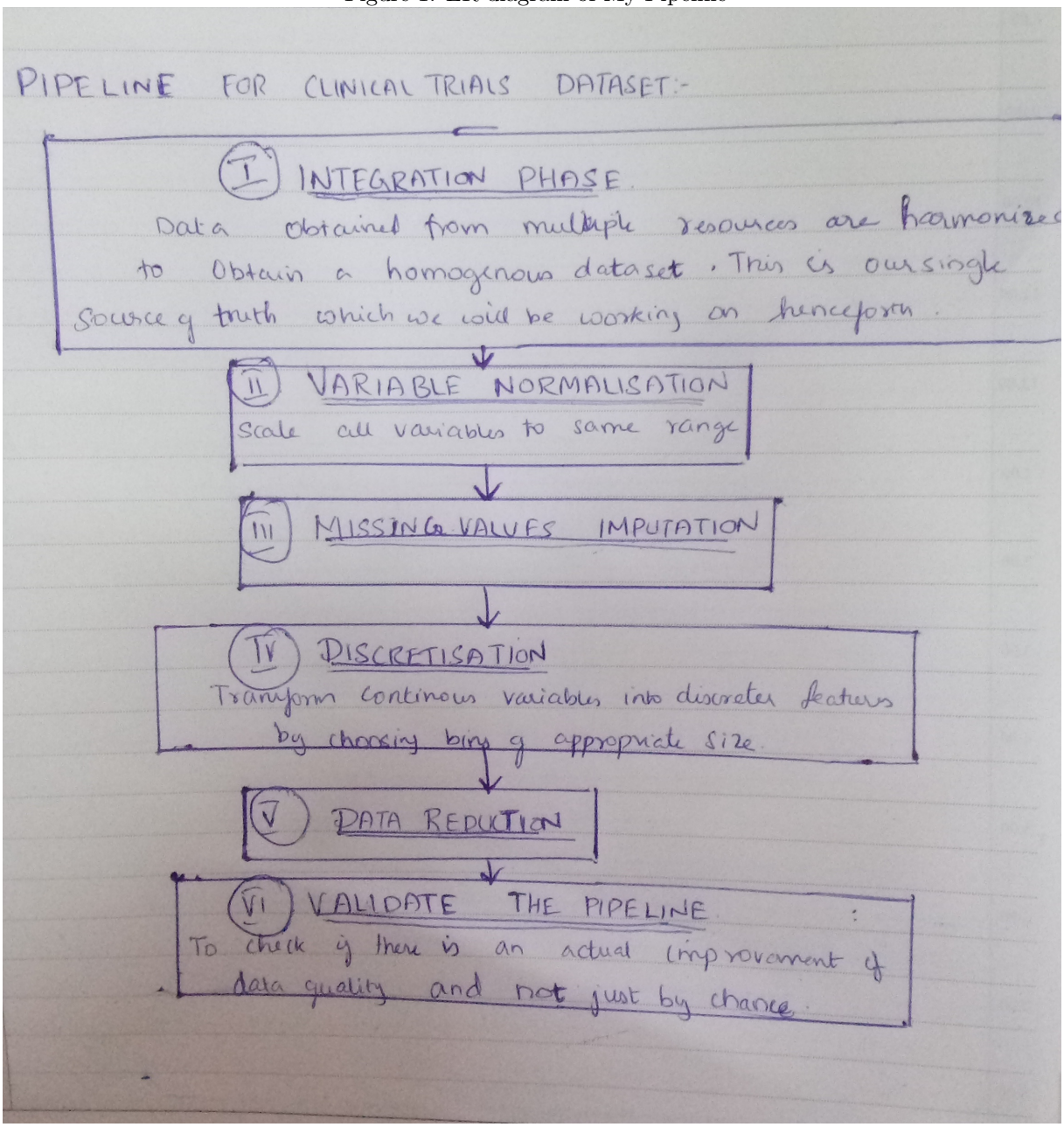
2.1 Problems faced:

- Inconsistencies when dealing with data coming from multiple resources (like same varibles reported in different measurement units or same varibles reported with different names etc)
- The varibles of dataset might be in different ranges.Many machine learning algorithms don't scale well when the variables are presenting very large and different ranges .
- Noise in data/missing values
- Redundant/irrelevant variables

2.2 Methods for solving above problems

- Data Harmonization :To solve for the first problem stated above ,we introduce this component in the pipeline
- Data Transformation : To solve for the second problem stated above,we normalise/standardise all variables and if required ,we dicretize some continous variables
- Data Cleaning : To check for outliers,anomalies and noise in data and remove them
- Having useless variables not only affects model performance but also the training time and introduce unecessary complexity in the model.

Figure 1: ER-diagram of My Pipeline



3 Step:1 Data Harmonization

When dealing with clinical trials dataset,the instruments used to measure certain variables might differ from place to place in terms of measurement units,as well as measurement precision.To make up for this ,we Bring them all to one single unit or even better ,we can make them dimensionless quantity.How do we do this?We measure the ratio of relative change from a particular threshold rather than using the absolute values.

$$newvar = \frac{oldvar - threshold}{threshold}$$

4 Varible Normalisation:

As discussed already,it is always desiarble to have all varibles scaled to similar range.To do this ,we might either use normalisation or standardisation.Between these two, we use standardisation if there are too many outliers or normalise other-wise.But normalisation does not guarantee same range for all variables.They must be used with dicretion.

5 Missing Value Imputation:

When it comes to clinicaltrials data,it is not uncommon for them to release only partial information about pipeline drugs and clinical trials to protect trade secrets or simply because there was no incentive to do more.But,deleting all observations with any missing factors greatly reduces the amount of data available and decreases the statistical power of the resulting statistics.To find a way around this,4 methods of imputation are proposed below.

5.1 Methods:

- Method1 : Unconditional mean imputation
- Method2 : K-nearest neighbor (kNN) imputation
- Method3 : Multiple Imputation
- Method4 : Decision-tree algorithms to impute missing variables

5.2 How they work:

- Unconditional mean imputation involves replacing the missing values for an individual variable with it's overall estimated mean from the available cases.Downside is that reduces data variane due to substuting more points at mean.
- KNN imputation involves selecting the distance measure in accordance with the type of varible dealt with (eg eucledian for nuerical and manhattan distance for categorical) the k hyperparameter.
- MI is a slighly more sophisticated imputer and i dont know much about it
- Decision tree algorithm : In order to do this,the feature containing missing values might need to be discretized,then if we apply our decision tree model upon them; class of missing values can be obtained

6 Discretization:

This component involves splitting the range of a variable into bins and create discrete class variables in place of continous variables.

6.1 Why discretize?

- to reduce noise due to small variations of values
- to decrease the amount of values to be memorized and managed
- to improve the classification performances

7 Data Reduction

This is probably one of the most important steps of the pipeline process,selecting important features in other words called dimensianlity reduction.This can done using follwing methods:

7.1 Methods:

- Missing Values Ratio
- Forward Selection
- Backward Selection
- stepwise selection
- Principal component analysis

8 Pipeline Validation

Alright,we have built a pipeline making many significant changes to the dataset.But what is the guarantee that our transformed dataset is of more high quality than the orginal one?We have got to validate our pipeline much like we validate a model.To validate the proposed pipeline and to prove the improvement of data quality, we assess the capability of the datasets obtained after each step to make predictions better than the un-transformed dataset