IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

HANDLING MISSING DATA TO IMPROVE GENERALIZATIONPERFORMANCEO MACHINELEARNING CLASSIFIER

R.Savithiri.[1,]
C.Kalaiarasi[2] K. Varalakshmi [3] A.Vijayanarayanan[4]
1, 2, 3, 4, Assistant Professor COMPUTER SCIENCE AND ENGINEERING
PERI INSTITUTE OF TECHNOLOGY

Abstract

In supervised learning, missing values usually appear in the training set. The missing values in a dataset maygenerate bias, affecting the quality of the supervised learning process or the performance of classification algorithms. These imply that a reliable method for dealing with missing values isnecessary. In this project, we analyze the difference betweeniterative imputation of missing values and single imputation inreal-world applications. We propose an iterative imputationmethod, in which each missing attribute-value is iterativelyfilled using a predictor constructed from the known values and predicted values of the missing attribute-values from the previous iterations. Meanwhile, we demonstrate that it is reasonable to consider the imputation ordering for patching upmultiple missing attribute values, and therefore introduce amethod for imputation ordering. We experimentally show that our approach significantly out performs some standard machine learning methods for handling missing values in classification tasks.

Introduction

The missing data problem is arguably the most common issue encountered by machine learning practitioners when analyzing real-world data. In many applications ranging fromgene expression in computational biology to survey responsesin social sciences, missing data is present to various degrees.As many statistical models and machine learning algorithmsrely on complete data sets, it is key to handle the missing data appropriately. Missing data problem is a common issue in most real-world studies. Since most statistical models and data-dependent machine learning (ML) algorithms could only handle complete datasets, the issue of how to approach missing values plays an important role in statistical inferences.Let Y be an (N × K) data matrix with i-th row yi = (yi1, yi2, ...,yiK) where yij is the value of j-th feature for the i- the sample. Define the subset of observed values as Y obs and missing values as Y mis. Also, let M = [mij] be the missing indicator matrix, where mij indicates whether yij is missing .ornot.Rubin(19776)defines three different missing mechanisms according to the conditional probability of the missing ,{mij = 1}, given the data. The mechanism of missing data is completely at random (MCAR) if the probability of missing ness is

independent of all datavalues,missing or observed, P(mij = 0|Y) = g(ϕ), i = 1, ..., N, j = 1, ..., K, whereg(.) is a known link function and ϕ is the vector of unknown mechanism parameters.The missing mechanism is called missing at random (MAR) if the probability of missingness depends only on the observed data values, P(mij = 0|Y) = g(Yobs; ϕ), i = 1, ..., N, j = 1, ..., K. Finally, the mechanism iscalled missing not at random (MNAR) when the probability of missingness may also depend on the unobserved data evenafterconditioningontheobservedones.Themissingmechanism for the likelihood inferences is ignorable when the MCAR or MAR assumptions hold with the additional condition of dis joint parameter spaces of the missing mechanism and the data model (see Little and Rubin, 2014; Tsiatis, 2007, for more details).One simple approach to analyze incomplete data is complete case (CC) analysis whichdiscards all incomplete cases. This approach is logical only if the missing rate is considerably small or the missing data mechanism is MCAR (Little and Rubin, 2014). However, ifthe missing mechanism is MAR or MNAR or the missing rateis considerably high, the CC approach could highly influence statistical results. This is due to the fact that CC analysis makes no use of observed features of an in complete case.

SYSTEM ANAYSIS EXISTING SYSTEM

In data mining process the biggest task of data preprocessing is missing value imputation .Imputation is a statistical process of replacing missing data with substituted values. Many clinical diagnostic dataset are usually incomplete. Excluding incomplete dataset from the originaldataset can bring more problem than simplification. In this paper the machine learning techniques for missing value imputation have been explored using Ionosphere data from UCI repository. The data imputation problem has been approached using well-know machine learning techniques.

DEMERITS

Some algorithms cannot handle missing values properly,while some techniques give efficient results to estimate themissing values. It is very important to handle missing data because many machine learning algorithm performances reduces due to missing values.

PROPOSED SYSTEM

In the dataset, there are few missing values(yet found to be hyper parameter),and pre-processing with such missing values is a common yet challenging problem. Re substitution will give biased results from the data to be observed for HD diagnosis and will certainly affect the valueof the learning process in Machine Learning In this approach,we impute each missing data attribute value by predicting its data value from non- missing data attributes. The experiments are conducted on benchmark medical datasets missing valuesranging from 1.98% to 50.65% and compared with iterativeimputation.

MERITS

In this approach, we impute each missing data attribute value by predicting its data value from non- missing data attributes. The experiments are conducted on benchmark medical datasets missing values ranging from1.98% to50.65% and compared with iterative imputation.

Data Sets

A. Dataset Obesity With Attributes and Class

Estimation of obesity levels based on eating habits and physical condition Data Set. Dataset include data for the estimation of obesity levels in individuals from the countries of Mexico, Peru and Colombia, based on their eating habits and physical condition. This data contains 17 attributes and 2111 records.

B. Imputation & Mice Implementation

After the missing value has been successfully created on purpose, the next step is to substitute the missing value with the imputation technique. This project will perform imputation with 2 techniques, which are Statistical and MICE .Done by using the fancy impute library , two ways to impute :KNN or K-Nearest Neighbor MICE or Multiple Imputation by Chained Equation for fancy impute : https:/ /pypi.org/ project/ fancyimpute / . In fancy impute, the MICE algorithm is named as Iterative Imputer. The steps taken are: Copy data_mv to obesity_mice_mpute Initialize Iterative Imputer with name mice_imputer Impute using fit_transform on data

C. Correlation Coefficient

The correlation coefficient is a statistical measure of thestrength of the relationship between the relative movements of two variables.The values range between -1.0 and 1.0.A correlation of-1.0 shows a perfect negative correlation, while a correlation of 1.0 shows a perfect positive correlation. Acorrelation of 0.0 shows no linear relationship between the movements of the two variables. variable value (y) based on agiven independent variable (x). Hence, the name is Linear Regression.

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

It has two types:

- Simple Linear Regression

If a single independent variable is used to predict the value of a numerical dependent variable, then such a Linear Regression algorithm is called Simple Linear Regression.

- Multiple Linear regressions

If more than one independent variable is used to predict thevalue of a numerical dependent variable, then such a Linear Regression algorithm is called Multiple Linear Regression.

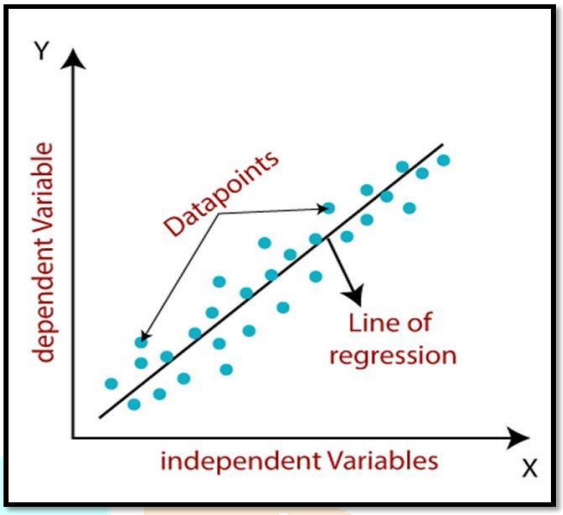

Fig 1: Regression

B. Logistic regression Algorithm

Logistic regression is basically a supervised classification algorithm. In a classification problem, the target variable(or output), y, can take only discrete values for a givenset of features(or inputs), X. The model builds a regressionmodelpredict the probability that a given data entry belongsto the categorynumberedas“1”.

MICE ALGORITHM

A. LINEAR REGRESSION ALGORITHM

Linear Regression is a machine learning algorithm based on supervised learning. It performs a regression task .Linear regression performs the task to predict a dependent

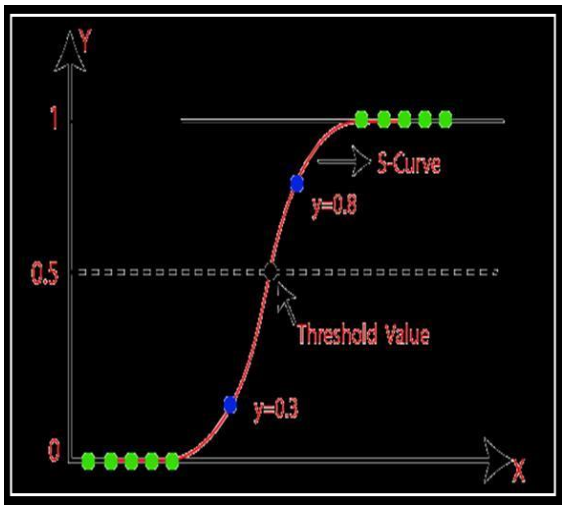

Fig 2: Linear Regression

C. RIDGE CLASSIFIER ALGORITHAM

A Ridge regressor is basically a regularized version Linear Repressor. i.e. to the original cost function of linear repressor we add a regularized term that forces the learningalgorithm to fit the data and helps to keep the weights lower a possible. The regularized term has the parameter,, alpha "which controls the regularization of the model i.e helps in reducing the variance of the estimates. Cost Function for Ridge Regressor.

$$J(\Theta) = \frac{1}{m}(X\Theta - Y)^2 + \alpha\frac{1}{2}(\Theta)^2$$

CONCLUSION

The choice of missing handling methodology has a significant impact on the clinical interpretation of the accompanying statistic a analyses. With missing data, the choice of whether to impute or not, and choice of imputationmethod, can influence the clinical conclusion drawn from aregression model. We recommend researchers to perform asensitivity analysis including at least MICE method. Often thecommon methods like mean, median ,mode, frequent data and constant doesn't provide the correct data for the missing values. The model is only as good as the data, so having acomplete dataset with proper data is a must; consider using MICE algorithm when you need to impute missing data.

FUTURE ENHANCEMENT

Missing data imputation using MICE was successfully performed on Missing at Random data using the fancy impute library. The results show that MICE performance is better than Statistical Imputation in terms of MSE and RMSE for numeric data , and Accuracy for categorical data. The MICE idea of using information from other columns proved helpful in replacing missing values. MICE is able to perform well on numeric and categorical data types. It is believed that MICE will produce the more excellent imputation when the correlation of each column is greater.

REFERENCES

Little

[1] Little & Rubin; "Missing data and imputation methods"PaperonMeanImputation1987.

[2] ZoubinGhahramani and Michael I. Jordan ; "Supervisedlearning fromincompletedataviaanEMapproach"1994.

[3] VanBuuren,Stef Groothu is-Oudshoorn,Karin" Mice: Multivariate Imputation by Chained Equations in R"2011.

[4] A. P. Dempster, N. M. Laird and D. B. Rubin "Maximum Likelihood from Incomplete Data via

theEMAlgorithm" 1977.

[5] Little, R.J.A. and Rubin, D.B. (1987) Statistical AnalysiswithMissingData.JohnWiley& Sons, NewYork.

[6] Y.Burgette and Reiter "Multiple Imputation: A Review ofPracticalandTheoreticalFindings" 2010.

[7] Daniel J. Stekhoven and Peter Bühlmann "Miss Forest—non-parametric missing value imputation formixed-type data"2010

[8] Dimitris Bertsimas &Bart Van Parys;"SparseHigh-Dimensiona lRegression: Exact Scalable Algorithmsand Phase Transitions" 2014

[9] DimitrisBertsimas&JackDunn"Optimalclassificationtrees" 2017

[10] ZhangEtAl&CaiEtAl;"FromPredictiveMethodstoMissingDataImputation:AnOptimizationApproach"2018.