STATISTICAL ANALYSIS WITH MISSING DATA

Yana Bondarenko

Oles Honchar Dnipro National University, Ukraine
e-mail: yana.bondarenko@pm.me

Abstract

Research of advanced techniques for processing multidimensional missing data is presented. The theoretical part of study is focused on the review of the different data imputation methods to handle missing data. The practical part of study is presented machine learning algorithms such as random forest, logistic regression, and nearest neighbor method to solve classification problem for data with artificial and natural missing values.

Keywords: missing value, imputation method, machine learning algorithm, accuracy

Problem formulation. Data matrix $X^{(n\times d)}$ is specified. Some of the values are not observed. It is necessary to restore data matrix for the application of machine learning algorithms to solve classification problems.

Removing objects with missing values. One of the simplest ways to solve problem of processing missing values is to remove the objects (rows of data matrix) that have missing values. This method is used when a small number of objects is not observed. The loss of information is the only drawback of this method. Removing features that have a small number of missing values is an alternative method.

Imputation with a special value. The second easiest method is to impute the missing data with a special predefined value (for example, zero or minus one). This approach does not reduce the sample size, but it can add to data matrix values that are significantly different from the real ones. It is appropriate to impute missing values with a value that does not occur in the data matrix (for example, minus one for non-negative values) for further application of machine learning algorithms based on decision trees and impute missing values with zero for algorithms that are sensitive to the features scale.

Mode imputation. The third easiest method is to impute the missing values of categorical feature by mode of the non-missing values of that feature.

Mean imputation. The fourth simplest method is to impute the missing values of numerical feature by mean of the non-missing values of that feature.

SVD imputation. Let's consider the method of replacing missing values using the singular value decomposition of data matrix. First, missing values are replaced by mean of the non-missing values for each feature in data matrix, then, these mean values are replaced by the nearest unique values for each feature to preserve the nature of data.

Second, the decomposition of data matrix is found in the form

$$X^{(n\times d)} = U^{(n\times n)} S^{(n\times d)} V^{(d\times d)},$$

where $S^{(n\times d)}$ is a singular matrix (it is diagonal matrix with the roots of the eigenvalues of the matrix $X^{(n\times d)}\left(X^{(n\times d)}\right)^T$ in descending order on the main diagonal). The matrices $U^{(n\times n)}$, $V^{(d\times d)}$ are orthogonal and, in addition, the columns of the matrix $U^{(n\times n)}$ are the eigenvectors of the matrix $X^{(n\times d)}$, and the matrix $V^{(d\times d)}$ can be presented in the form $V^{(d\times d)} = \left(S^{(n\times d)}\right)^{-1}\left(U^{(n\times n)}\right)^T X^{(n\times d)}$.

Third, the first r rows and columns are selected in matrix $S^{(n\times d)}$, and all remaining ones are deleted. The first r most significant singular values are called principal components.

Fourth, data matrix can be restored having selected the first r columns in the matrix $U^{(n\times n)}$, and the first r rows in the matrix $V^{(d\times d)}$:

$$X_{approx}^{(n\times d)} = U^{(n\times r)} S^{(r\times r)} V^{(r\times d)}.$$

Fifth and finally, the values in the places of missing values in matrix $X^{(n\times d)}$ are replaced on the values obtained in the reconstructed matrix $X_{approx}^{(n\times d)}$.

Steps 2, 3, 4 can be repeated for a predefined number of iterations to improve the recovery of the data matrix or to use the quality criteria of matrix recovery by calculating the proximity to 1 for the coefficient of determination:

$$Q(r) = \frac{\sum_{k=1}^{r} \lambda_k}{\sum_{k=1}^{n} \lambda_k},$$

where λ_k are the eigenvalues of the matrix $X^{(n\times d)}\left(X^{(n\times d)}\right)^{\mathrm{T}}$. The dependence of the coefficient of determination $Q(r)$ on the number of principal components r allows to evaluate the efficiency of the method. At the end, the reconstructed values in the places of missing values in matrix $X^{(n\times d)}$ are replaced by the nearest unique values for each feature of matrix $X_{approx}^{(n\times d)}$ to preserve the nature of data.

Nearest neighbor imputation algorithm. A hypothesis about similar values of features for close objects is proposed. Thus, the missing values of the features for a certain object can be restored using the known values of the features of k nearest neighbors of this object. Let's consider the method of replacing missing values using the nearest neighbor imputation algorithm.

First, the mask of data matrix $X^{(n\times d)}$ from the Boolean variable True (missing value) and False (no missing value) is created. And this mask is applied to find the number of missing values in each object of the data matrix.

Second, the mask of objects from Boolean variables True (missing value in features) and False (no missing value in features) is created. And this mask is applied to creation of matrix of objects X^{full} with non-missing values.

Third, the mask of each object from Boolean variables True (missing values in features) and False (no missing values in features) is created. And this mask is applied to find features with non-missing values in each object in the data matrix.

Fourth, the distances between the objects of X^{full} and the object of $X^{(n\times d)}$ is calculated (it should be noted that the square of the distance between two objects is equal to the sum of the squared distances between them for each feature with non-missing values).

Fifth, the distances are sorted according to the ascending order and the k smallest are selected, hence the k nearest neighbors for each object are found.

Sixth, missing values are replaced by mean for each feature of the k nearest neighbors in each object in the data matrix.

At the end, the mean values in the places of missing values in matrix $X^{(n\times d)}$ are replaced by the nearest unique values for each feature of matrix $X^{(n\times d)}$ to preserve the nature of data.

It is appropriate to use algorithm if there is a large number of objects with non-missing values, otherwise, at first it is necessary to replace missing values by mean for each feature, and after that, the values in the places of missing values are replaced by the nearest unique values for each feature to preserve the nature of data, and finally, the entire data matrix is selected to be the matrix of objects X^{full} with non-missing values.

Random forest imputation algorithm. First, missing values are replaced by mean of the non-missing values for each feature in data matrix. Second, these mean values are replaced by the nearest unique values for each feature to preserve the nature of data. Third, prediction for each feature with missing values is made using the random forest algorithm (at the same time, training has been implemented on objects with non-missing values for this feature). Fourth, replacement of missing values in each feature is carried out using the prediction of decision trees composition obtained above. At the end, the values in the places of missing values in data matrix are replaced by the nearest unique values for each feature of matrix to preserve the nature of data.

Linear regression imputation algorithm. First, missing values are replaced by mean of the non-missing values for each feature in data matrix. Second, these mean values are replaced by the nearest unique values for each feature to preserve the nature of data. Third, prediction for each feature with missing values is made using linear regression algorithm (at the same time, training has been implemented on objects with non-missing values for this feature). Fourth, replacement of missing values in each feature is carried out using the prediction with linear regression. At the end, the values in the places of missing values in data matrix are replaced by the nearest unique values for each feature of matrix to preserve the nature of data.

k-means imputation algorithm. A hypothesis about similar values of features for close objects is proposed. Thus, the missing values of the features for certain object can be restored using the known values of the center of cluster, which owns the object with missing values.

Initial data and specific features of implementation methodology. Six different data sets were used to compare the quality of missing data replacement. Three data sets with complete values and three data sets with natural missing values were studied. Machine learning algorithms were applied to solve classification problems.

Complete data sets were used to estimate the performance of machine learning algorithms by selecting a different proportion of missing data. In addition, information about true values of artificial missing value allows to compare the recovered data directly. Missing data were created artificially for complete data according to the following scheme: 1) subset of one fourth of the most important features is selected with random forest algorithm (it should be reminded that random forest is able to estimate the importance of features based on the frequency of each feature during decision tree construction). This subset of features is used in all experiments; 2) missing value with a certain probability is created for each value from the subset of selected features so that the proportion of all missing values was the same as the given one.

Complete data sets (AI4I 2020 Predictive Maintenance Dataset Data Set, Banknote Authentication Data Set, Car Evaluation Data Set), as well as data sets with natural missing values (Cargo 2000 Freight Tracking and Tracing Data Set, Cervical cancer (Risk Factors) Data Set, HCC Survival Data Set) were selected from the UCI Machine Learning Repository.

The following parameters for imputation algorithms were used. Imputation with a special value: missing values were replaced with minus one when using a random forest algorithm, and missing values were replaced with zero when using a logistic regression algorithm. SVD imputation: the rank of the matrix $X_{approx}^{(n\times d)}$ is half the number of features, maximum number of iterations is 10. Nearest neighbor imputation algorithm: number of nearest neighbors is 5, space metric is L_2. Random forest imputation algorithm: number of decision trees is 10, maximum number of iterations is 3. Linear regression imputation algorithm: maximum number of iterations is 3. k-means imputation algorithm: number of clusters is 8, maximum number of iterations is 3.

Results. Data sets with a different proportion (from 2,5% to 15% with a step 2,5%) of missing data were created in one fourth of the most important features selected with random forest algorithm. Accuracy and RMSE were calculated according to 10-fold stratified cross-validation sample.

Dependences of the classification accuracy of recovered data on the proportion of missing values are presented in Figure 1a, Figure 1b, Figure 1c.

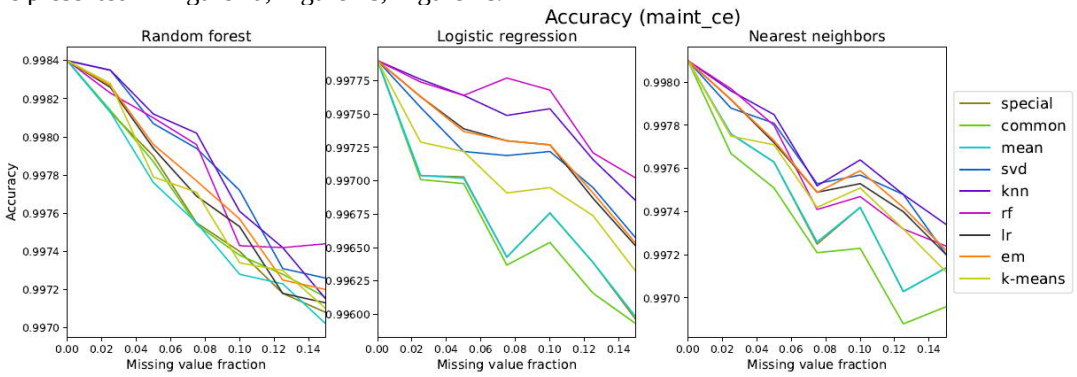

Fig. 1a. Dependence of the classification accuracy of recovered data
on the proportion of missing values (AI4I 2020 Predictive Maintenance Dataset Data Set)

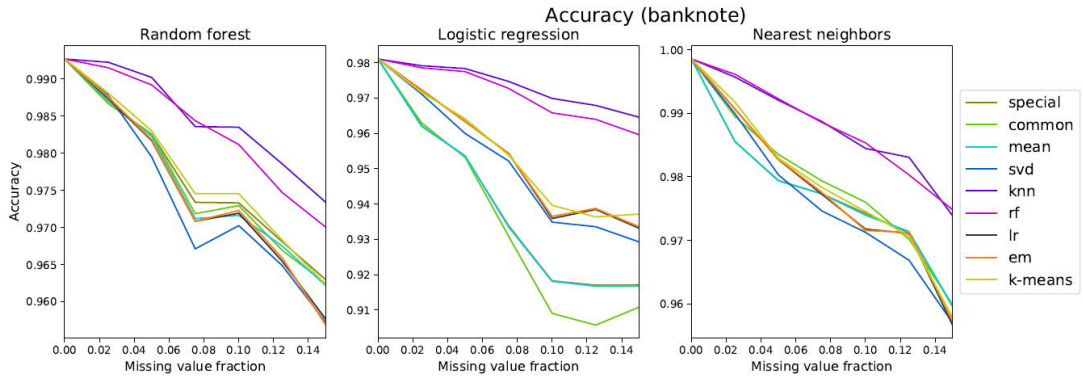

Fig. 1b. Dependence of the classification accuracy of recovered data on the proportion of missing values (Banknote Authentication Data Set)

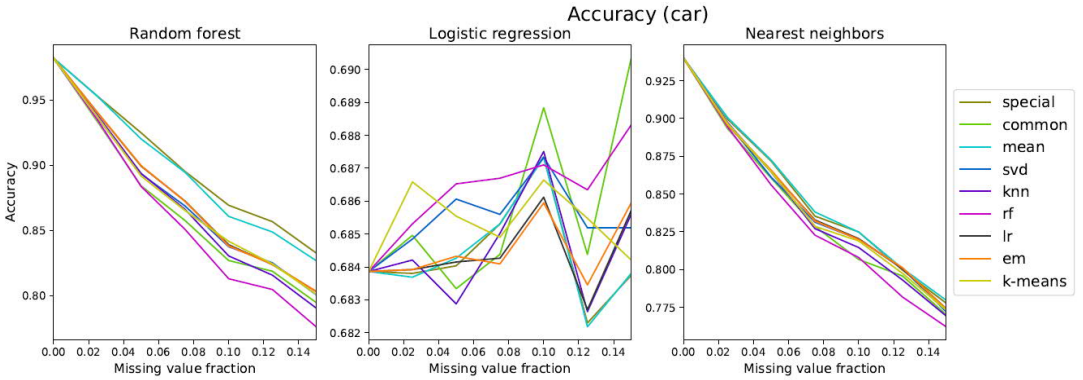

Fig. 1c. Dependence of the classification accuracy of recovered data on the proportion of missing values (Car Evaluation Data Set)

Dependences of the RMSE between recovered and real data on the proportion of missing values are presented in Figure 2.

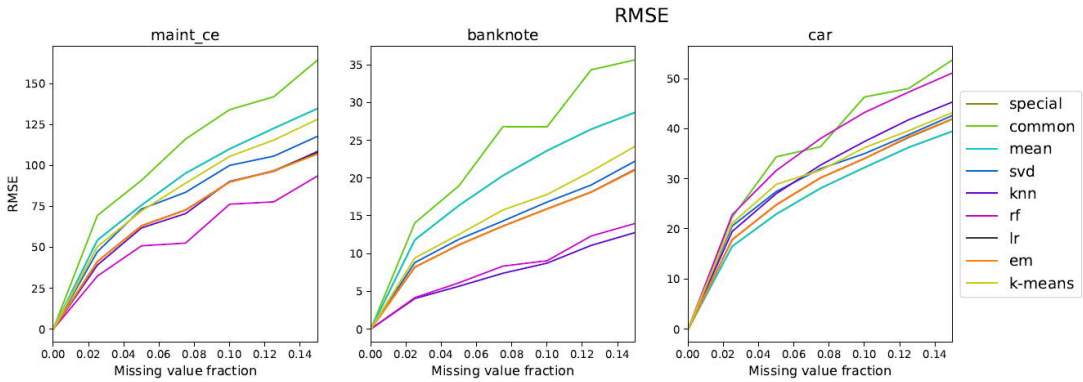

Fig. 2. Dependence of the RMSE between recovered and real data on the proportion of missing values

Performance outcomes for data with natural missing values are shown in Table 1. The best and closest results in each column are highlighted in bold.

Experiments have shown that there is no universal method for missing value replacement, which would be superior in accuracy to all others. Simple imputation methods (such as mode imputation,

mean imputation, imputation with a special value) have demonstrated performance comparable to advanced imputation methods (such as k nearest neighbors, random forest, linear regression, k means) in case of data with natural values.

Table 1. Classification accuracy for data sets with natural missing values

Datasets	Cargo 2000			Cervical cancer			HCC Survival		
Methods	RF	LR	KNN	RF	LR	KNN	RF	LR	KNN
special	**0.9997**	0.3555	0.5630	0.9953	0.9918	0.9918	0.7327	**0.7580**	**0.6669**
mean	**0.9997**	0.3555	0.5630	0.9953	0.9918	0.9918	0.6900	**0.7580**	**0.6669**
SVD	0.9903	0.6127	0.5653	**0.9965**	0.9918	0.9918	0.7323	0.7040	0.6415
KNN	**0.9997**	0.5830	0.5721	0.9930	**0.9930**	0.9918	**0.7463**	0.7394	0.6724
RF	0.9741	0.7813	0.5546	0.9941	0.9918	0.9918	0.7084	0.6970	0.6591
LR	0.9974	0.8092	0.8488	0.9953	0.9918	0.9918	0.7029	0.7150	0.6661
k-means	0.9997	0.5526	0.5052	0.9964	0.9918	0.9918	0.7455	0.7518	0.6536

Selection of the imputation method may depend on the types of features with missing values, on the number of objects with missing values, and on the cause of missing values. Each problem requires an individual approach for imputation missing values.

References

Bache, K. and Lichman, M. (2013) UCI Machine Learning Repository. University of California, School of Information and Computer Science. http://archive.ics.uci.edu/ml

Bishop, C.M. (2007) *Pattern Recognition and Machine Learning*. Springer, New York.

Hastie, T., Tibshirani, R., and Friedman, J. (2009) *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, New York.

Kayumov E. (2016) Imputation methods for missing values. https://github.com/emilkayumov/missing-value

Little, R. J. A., Rubin, D. B (2002) *Statistical Analysis with Missing Data*. John Wiley & Sons, New York.

Murphy, K.P. (2012) *Machine Learning: A Probabilistic Perspective*. The MIT Press, Cambridge.

VanderPlas J. (2016) *Python Data Science Handbook. Essential Tools for Working with Data*. O'Reilly Media, Inc.