A New Method to Missing Value Imputation for Immunosignature Data

DOI: 10.17691/stm2019.11.2.03
Received May 3, 2018

© **A.A. Koshechkin,** Laboratory Assistant, Scientific and Educational Center of Computer Science and Technology;
V.S. Andryushchenko, Laboratory Assistant, Scientific and Educational Center of Computer Science and Technology; Programmer, Institute of Applied Mathematics and Computer Science;
A.V. Zamyatin, DSc, Head of the Department of Theoretical Foundations of Informatics; Director of the Scientific and Educational Center of Computer Science and Technology

National Research Tomsk State University, 36 Lenin Avenue, Tomsk, 634050, Russia

The immunosignature technology uses microarray chips of random amino acid sequence peptides to detect diseases based on the change in the profile of circulating antibodies. Diseases are detected using classification algorithms trained on a reduced sample of immunosignature patterns of patients with known diagnoses.

The aim of the study was to develop a new method of missing value imputation in immunosignature data, which allows maintaining sufficient accuracy of classification.

Materials and Methods. The study was carried out using immunosignature data obtained by utilizing a high-resolution peptide microarray chip with nearly ten thousand peptide cells.

The applicability of various missing value imputation methods such as simple imputation, weighted k-nearest neighbors and machine learning techniques (linear regression, random forest, gradient boosting) was evaluated.

Results. Missing value imputation method based on gradient boosting has been developed in the framework of the study. Its operating principle implies iterating through all features (attributes) and training on examples (samples) whose values are present in the feature, followed by clarification of missing feature values. This process is repeated until the total training error for all features stops decreasing or until the maximum number of iterations is reached. The root mean squared error is employed as a training error metric.

To assess the quality of missing value imputation, classification results based on the data obtained after imputation procedure are used in our investigation.

The proposed missing value imputation algorithm based on linear gradient boosting proves to be effective under conditions of a high proportion of missing values as compared to other methods under consideration. The results of the investigation demonstrate the viability of using machine learning techniques for missing value imputation in immunosignature data.

Key words: early diagnosis of diseases; immunosignature; missing value imputation in the sample; machine learning.

Introduction

Today, fast development of modern technology makes large amounts of various information available for analysis. However, there is often a problem of data representativeness unavoidable in collecting and analyzing this information. Representativeness declines mainly due to missing values under the influence of noise and the human factor when data are collected. Missing values have become a challenging issue for researchers, since all data mining methods work incorrectly or become ineffective in such situation. Therefore, missing value imputation in the sample is one of the primary tasks in data mining [1].

Currently, many cancers are potentially curable if diagnosed in an early stage. Early detection of malignant tumors requires simple, inexpensive, minimally invasive and accurate diagnostic methods [2]. The immunosignature technology is one of such methods [3].

This technology employs microarray chips of random amino acid sequence peptides to detect diseases based on the change in the profile of circulating antibodies [4]. These peptides have partial or complete similarity to antigen epitopes. Using the arrays of peptides that represent probable amino acid sequences of proteins makes it possible to determine the binding partner for many antibodies, even if there is no exact match for the epitope [5]. The data obtained are expected to help health care professionals make the final diagnosis. However, errors may occur in the process of obtaining the data due to both technological properties of the equipment and subsequent data digitization. Data representativeness may be lost because of this

Corresponding author: Alexander A. Koshechkin, e-mail: kaa1994g@mail.ru

and some peptide luminosity values may become unavailable. This explains the urgency of solving the problem of adequate missing value imputation.

The aim of the study was to evaluate the existing methods for missing value imputation in the selected dataset and develop a new efficient method for processing such values.

To achieve this goal, it was necessary to solve the following tasks:

to evaluate the existing missing value imputation methods based on various mathematical tools;

based on the obtained knowledge, to offer the most suitable method for missing value imputation, applicable to immunosignature data;

to compare efficiency of the existing and proposed methods for missing value imputation in immunosignature data.

Materials and Methods

Materials obtained by immunosignature assay are a set of fluorescence intensity values for peptides, where the peptide names are columns (features), and class labels (samples) are rows. A set of data from the public repository of biomedical data (accession GSE52580) obtained by digitizing immunosignature data was used in the study [6, 7].

The dataset has no missing values, which allows us to control the nature and number of missing values in our set. The dataset has the following characteristics:

the number of samples — 240;

the number of features — 9781;

the number of classes — 6.

The number of samples is the same in each class, so the dataset is a balanced subset. Table 1 shows a fragment of a dataset with known disease classes and peptides. The names of peptides are presented as sequences of amino acids.

Missing value imputation methods under study. To date, there are a large number of different missing value imputation methods applicable to the solution of this problem [8–10], therefore it is reasonable to analyze methods based on fundamentally different mathematical tools. The following selected methods have been considered in more detail.

Table 1

Fragment of the peptide matrix

Disease	Peptide name	
	CSGYNSFAMKANYIFNG	CSGSNYYDWWFRIAVMITI
Brain cancer	5.27889752	9.15952333
Breast cancer	0.89180777	0.89329176
Esophageal cancer	0.88392227	1.12217693
Multiple myeloma	0.82533253	0.93682348
Pancreatic cancer	0.96485786	1.02698893
Healthy control (healthy individuals)	0.85648045	0.84041385

Simple imputation is one of the simplest and most well-known missing value imputation methods [11]. It consists of replacing the missing values of a feature with the median, mean or mode calculated from the present values of the feature. The advantage of this method is quick missing value imputation. However, if the quantity of these values is large, simple imputation will lead to significant distortion of data analysis results.

Weighted k-nearest neighbors algorithm is a simple and effective method to handle missing values based on the hypothesis stating that if the examples (samples) are close in the measured feature space, this implies they are close by unmeasured features [12]. The distance between two examples is calculated from the present feature values. Weighted average values of neighbors are used to calculate missing values [13].

Missing values can be recovered in a particular feature by predicting its values by other features using a variety of machine learning techniques. This approach implies successive presentation of each feature as a target variable, which is followed by training on examples having no missing values in the target variable and subsequent prediction of missing values of the target variable.

Since there are missing values among the features by which training is performed, they should be initially replaced by one of the simplest imputation methods, and then clarified using one of the machine learning techniques. Let us take a closer look at some of the most popular machine learning techniques.

Linear regression is a machine learning technique that involves constructing equations with a linear function of target feature dependence on one or more other features [14]. It is possible to construct the equation of dependence of one feature on another for two features Y_1 and Y_2:

$$Y_2 = aY_1 + b,$$

using the values known for each feature and impute the missing values using the resulting regression equation for the available values. This method is effective only if there is a certain level of linear dependence between the features.

Random forest is a machine learning technique based on a multitude of decision trees combined into an ensemble [15]. This algorithm is a universal solution and works effectively with both continuous and categorical features [16]. For categorical features, the predicted value is determined by majority voting of each individual tree in the ensemble. In turn, for numerical features, the predicted value is determined as the average between responses of each tree in the ensemble.

Proposed missing value imputation method. Missing value

imputation method developed in the framework of this investigation is based on the use of such machine learning technique as gradient boosting.

Gradient boosting is a machine learning technique based on creating a linear combination of simple algorithms by changing the weight of the input data [17, 18]. Each simple algorithm (linear classifier or a decision tree) is created in such a way as to give more weight and preference to previously incorrectly predicted values. Linear combination length in simple algorithms is equal to the number of model rounds.

Since gradient boosting provides the possibility to build a linear combination of different algorithms, we can actually use two different methods — *linear gradient boosting and gradient tree boosting*.

To handle missing values, gradient boosting is used as follows. Missing values in the entire data set are imputed using simple imputation (in this paper, the missing values in each feature are replaced by a median calculated from the present feature values). Next, training is performed successively for each feature on samples whose values are present in this feature, followed by clarification of missing feature values. This process is repeated until the total training error for all features stops decreasing or until the maximum number of iterations is reached. The root mean squared error is employed as a training error metric [19].

Comparing the efficiency of missing value imputation methods. The main purpose of using the immunosignature technology is to support decision-making in diagnosis, which in the terms of data mining is reduced to the problem of classification. In this regard, to assess the quality of missing value imputation, classification results based on the data obtained by imputation procedure were applied in our investigation.

This data set was already applied in earlier studies [20] where random forest algorithm was found to show high classification results, therefore, using it in work as a classifier was quite appropriate. Various metrics are used to assess the accuracy of classification. Their selection and analysis are an indispensable part of any investigation [21]. Given that the dataset was a balanced selected subset, we applied a classification metric "proportion of correct answers" (accuracy) in our investigation.

The data set had a large number of features, which negatively affected the time spent on calculations. Therefore, to save time, only 120 most informative features were selected from this set using Student's t-test [22].

To compare the effectiveness of methods under consideration, it was necessary to create a subset with a wide range of parameters. It was appropriate to implement this by creating missing values artificially. In multi-step analysis of the peptide microarray, missing values may appear at any stage with no consistent pattern. Therefore, it was impossible to find a method for creating missing values that imitated all possible scenarios. In this regard, the method of completely random creation of missing values (missing completely at random) was used in this study [23, 24].

Table 2

Packages used and their parameters

Methods	Package/profile
Simple imputation	caret/medianImpute
Weighted k-nearest neighbors algorithm	wNNSel
Linear regression	ice/norm.predict
Random forest	missForest/100 trees, maximum iterations: 10
Linear gradient boosting	xgboost/maximum iterations: 10, maximum rounds: 100, eta = 0.3
Gradient tree boosting	

The methods were compared following these steps:

1. Preparing multiple datasets with different numbers of missing values.
2. Imputation of missing values in created datasets using each of the methods under consideration in turn.
3. Classifying based on the data of each imputed set.
4. Repeating steps 1–3 thirty times.
5. Calculating the average accuracy of classification carried out using the imputed datasets for each of the methods under consideration for different numbers of missing values.
6. Evaluating the results.

The work was performed using the R programming language and software libraries available in CRAN repository (Table 2).

Results

It is important to note that the investigation has revealed inability of such methods as linear regression and k-weighted nearest neighbors to process data sets, because the proportion of missing values in these cases exceeds 0.7 and 0.85, respectively. Therefore, performance assessment values are incomplete for these methods.

The results make it obvious (Figure 1) that the use of machine learning techniques to impute missing values is the most effective solution. Random forest and linear gradient boosting have shown the best results.

Table 3 presents the maximum and minimum accuracy values. At a proportion of missing values approximating 0.7, such methods as random forest, linear gradient boosting, and gradient tree boosting maintain the same variance indices, which indicates their high efficiency.

The data given in Table 3 and Figure 1 illustrate sharp worsening of the results for random forest and gradient tree boosting in datasets with a high proportion of missing values. Consequently, linear gradient boosting is more preferable in this situation. At the same time,

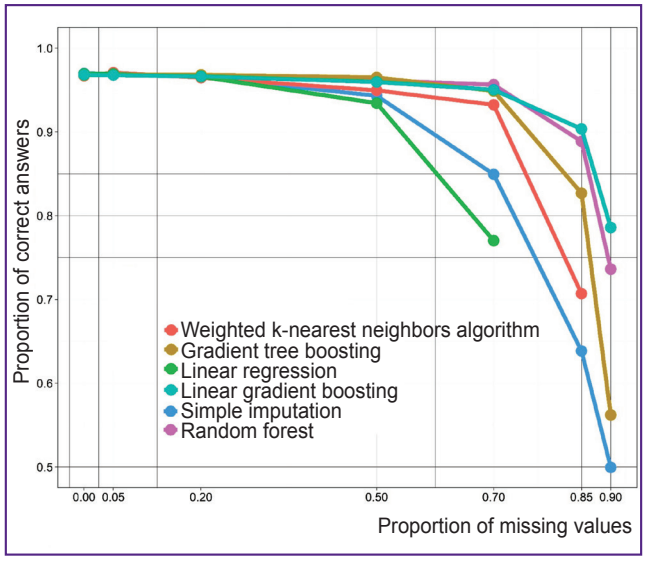

Figure 1. Classification results based on the data obtained after imputation of a variable proportion of missing values

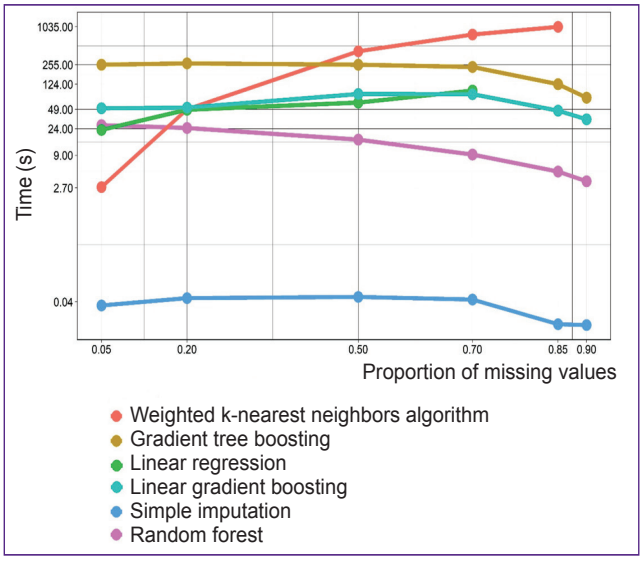

Figure 2. Processing time for methods with different proportions of missing values on a log scale

Table 3

Minimum and maximum values of classification results

Methods and models		Proportion of missing values					
		0.05	0.20	0.50	0.70	0.85	0.90
Simple imputation	max	0.974	0.973	0.956	0.872	0.668	0.553
	min	0.965	0.957	0.930	0.822	0.608	0.432
Weighted k-nearest neighbors algorithm	max	0.977	0.971	0.961	0.951	0.743	NA
	min	0.966	0.955	0.937	0.909	0.657	NA
Linear regression	max	0.974	0.971	0.947	0.806	NA	NA
	min	0.961	0.960	0.922	0.721	NA	NA
Random forest	max	0.973	0.973	0.971	0.972	0.921	0.811
	min	0.964	0.960	0.952	0.938	0.846	0.621
Linear gradient boosting	max	0.974	0.972	0.970	0.960	0.930	0.823
	min	0.963	0.957	0.950	0.931	0.877	0.691
Gradient tree boosting	max	0.975	0.975	0.974	0.973	0.861	0.694
	min	0.964	0.969	0.956	0.933	0.736	0.444

N o t e: NA (not available) — unavailable value.

there are no differences between the methods when the proportion of missing values is low.

Based on processing time values (Figure 2), two aspects can be emphasized in the methods.

Firstly, processing time of missing value imputation algorithms based on machine learning techniques decreases with the increase in proportion of missing values. This is due to the fact that the size of the training sample reduces with increase in the number of missing feature values and so does the training time of the algorithm.

Secondly, the method of k-weighted nearest neighbors is preferable for a small proportion of missing values as less time is required for missing value imputation.

Conclusion

The results of the investigation have revealed efficiency of the proposed method for missing value imputation based on linear gradient boosting under conditions of high proportion of missing values as compared to the analogous methods considered. At the same time, the method of k-weighted nearest neighbors is preferable for small numbers of missing values due to insignificant amount of time required for data processing and performance efficiency comparable to more complex methods.

The obtained knowledge is the basis for future research and eventual creation of software packages for pre-processing of peptide microarray data.

Study funding and conflict of interests. This study was not supported by any financial sources and the authors have no conflict of interests to disclose.

References

1. Padgett C.R., Skilbeck C.E., Summers M.J. Missing data: the importance and impact of missing data from clinical research. *Brain Impairment* 2014; 15(01): 1–9, https://doi.org/10.1017/brimp.2014.2.

2. Osipova T.V., Ryabykh T.P., Baryshnikov A.Yu. Diagnostic microchips: application in oncology. *Rossijskij bioterapevticeskij zurnal* 2006; 5(3): 72–81.

3. O'Donnell B., Maurer A., Papandreou-Suppappola A., Stafford P. Time-frequency analysis of peptide microarray data: application to brain cancer immunosignatures. *Cancer Inform* 2015; 14(Suppl 2): 219–233, https://doi.org/10.4137/cin.s17285.

4. Richer J., Johnston S.A., Stafford P. Epitope identification from fixed-complexity random-sequence peptide microarrays. *Mol Cell Proteomics* 2014; 14(1): 136–147, https://doi.org/10.1074/mcp.m114.043513.

5. Kukreja M., Johnston S.A., Stafford P. Immunosignaturing microarrays distinguish antibody profiles of related pancreatic diseases. *J Proteomics Bioinform* 2013; S6: 001, https://doi.org/10.4172/jpb.s6-001.

6. Stafford P., Halperin R., Legutki J.B., Magee D.M., Galgiani J., Johnston S.A. Physical characterization of the "immunosignaturing effect". *Mol Cell Proteomics* 2012; 11(4): M111.011593, https://doi.org/10.1074/mcp.m111.011593.

7. *National Center for Biotechnology Information Search database.* URL: https://www.ncbi.nlm.nih.gov/.

8. Efromovich S. Nonparametric regression with missing data. *Wiley Interdisciplinary Reviews: Computational Statistics* 2014; 6(4): 265–275, https://doi.org/10.1002/wics.1303.

9. Van Buuren S. Multiple imputation of discrete and continuous data by fully conditional specification. *Statistical Methods in Medical Research* 2007; 16(3): 219–242, https://doi.org/10.1177/0962280206074463.

10. Zloba E., Yatskiv I. Statistical methods of reproducing of missed data. *Computer Modelling & New Technologies* 2002; 6(1): 51–61.

11. Žliobaitė I., Hollmén J. Optimizing regression models for data streams with missing values. *Machine Learning* 2014; 99(1): 47–73, https://doi.org/10.1007/s10994-014-5450-3.

12. Troyanskaya O., Cantor M., Sherlock G., Brown P., Hastie T., Tibshirani R., Botstein D., Altman R.B. Missing value estimation methods for DNA microarrays. *Bioinformatics* 2001; 17(6): 520–525, https://doi.org/10.1093/bioinformatics/17.6.520.

13. Tutz G., Ramzan S. Improved methods for the imputation of missing data by nearest neighbor methods. *Computational Statistics & Data Analysis* 2015; 90: 84–99, https://doi.org/10.1016/j.csda.2015.04.009.

14. Little R.J.A. Regression with missing X's: a review. *J Am Stat Assoc* 1992; 87(420): 1227–1237, https://doi.org/10.2307/2290664.

15. Breiman L. Random forests. *Machine Learning* 2001; 45(1): 5–32.

16. Stekhoven D.J., Buhlmann P. MissForest — nonparametric missing value imputation for mixed-type data. *Bioinformatics* 2011; 28(1): 112–118, https://doi.org/10.1093/bioinformatics/btr597.

17. Natekin A., Knoll A. Gradient boosting machines, a tutorial. *Front Neurorobot* 2013; 7: 21, https://doi.org/10.3389/fnbot.2013.00021.

18. Friedman J.H. Greedy function approximation: a gradient boosting machine. *The Annals of Statistics* 2001; 29(5): 1189–1232, https://doi.org/10.1214/aos/1013203451.

19. Hyndman R.J., Koehler A.B. Another look at measures of forecast accuracy. *Int J Forecast* 2006; 22(4): 679–688, https://doi.org/10.1016/j.ijforecast.2006.03.001.

20. Andryushchenko V.S., Uglov A.S., Zamyatin A.V. Statistical classification of immunosignatures under significant reduction of the feature space dimensions for early diagnosis of diseases. *Sovremennye tehnologii v medicine* 2018; 10(3): 14–20, https://doi.org/10.17691/stm2018.10.3.2.

21. Arlot S., Celisse A. A survey of cross-validation procedures for model selection. *Statistics Surveys* 2010; 4: 40–79, https://doi.org/10.1214/09-ss054.

22. Student. The probable error of a mean. *Biometrika* 1908; 6(1): 1–25, https://doi.org/10.2307/2331554.

23. Rubin D.B. Inference and missing data. *Biometrika* 1976; 63(3): 581, https://doi.org/10.2307/2335739.

24. Tikhova G.P. Data missing: how to solve and how to escape the problem. *Regionarnaya anesteziya i lechenie ostroy boli* 2016; 10(3): 205–209.