Feature Engineering and Missing Data Imputation Method of Medical Data Analysis

Nataliya Shakhovska [1], Nataliia Melnykova[1]

[1] *Lviv Polytechnic National University, S.Bandera str,12, Lviv, 79013, Ukraine*

Abstract

This work provides an alternative way to preprocessing procedure for consolidated data. Two methods are proposed. The first one is used for feature selection based on ensemble of machine learning algorithms. And the second one organizes missing data imputation based on combination of functional dependencies and associative rules. Ensemble methods for processing multimodal data based on a hierarchical classifier, a set of weak classifiers and a number of methods for selecting important characteristics with a much higher value of accuracy on unbalanced data sets compared to existing machine learning methods are developed. The methods are validated on medical dataset. The percentage of recovery data is on 1.2% comparing with associative rules. The proposed missing data imputation method creates additional data values operating a based domain and functional dependencies and includes these values to available training data. The correctness of the filled-in values is proved on the predictor built on the original dataset. The proposed PPD method conducts 12% better than RF and EM models for 30% missing data.

Keywords

feature selection, missing data, machine learning, ensemble, data preprocessing.

1. Introduction

Methods for detecting dependencies in data have existed for more than a century. The first of them emerged as methods of mathematical and statistical analysis (correlation and factor analysis). The computerization of data analysis processes has dramatically increased the amount of data analyzed, the quality of analysis and the scope of methods of knowledge extraction.

The information boom has led to a thousand-fold growth in data collected in different domains. Such areas include computer technology, economics, sociology, medicine, astronomy, etc. The increase in the amount of information collected continues to grow exponentially. For example, based on a Digital Universe Study commissioned by EMC, the total global data in 2005 was 130 exabytes, up to 2227 EB by 2015, and increased again last year to 7 ZB (zettabytes). The forecast made by the same study shows that by 2025 the amount of digital data will increase to 10.9 ZB. The size of separate databases is growing just as fast and has overcome the petabyte barrier. Most of the collected data is not currently analyzed or is only a simple analysis. This is the most important for medical data, primarily when hospitals do not support medical standards for data exchanging (HL7, for example).

The main problems that arise in data processing are

- the absence of analytical methods appropriates for use in several subject areas,
- the demand for significant human resources to support the data investigation process,
- the high computational complexity of existing investigation algorithms and the fast growth of composed data.

To the constant increase in research time, even with frequent updates of server hardware and the need to work with distributed databases, the capacities of which most existing data analysis methods

CMIS-2022: The Fifth International Workshop on Computer Modeling and Intelligent Systems, Zaporizhzhia, Ukraine, May 12, 2022
EMAILnataliya.b.shakhovska@lpnu.ua (N. Shakhovska); melnykovanatalia@gmail.com (N. Melnykova)
ORCID: 0000-0002-6875-8534 (N. Shakhovska); 0000-0002-2114-3436 (N. Melnykova)

are not used inefficiently. Thus, there is a problem of developing an effective strategy of data analysis that can be applied to dispersed databases of various domains.

Data analysis methods are effectively used on clear and previously processed data. That is why the paper aimed to analyses and develop methods for data preprocessing. Particularly, we have taken into account feature selection (feature engineering) and missing data imputation.

The main contribution of the paper is as follows:

1. A new hybrid ensemble feature selection model for a machine learning-based post-COVID-19 prediction system is proposed as an automatic feature cut-off rank identifier.
2. A method based on probabilistic dependence is developed and tested. The percentage of recovery data is on 1.2% comparing with associative rules.
3. Development of a method for finding dependencies in large data sets with gaps and uncertainties based on an ensemble of clustering methods and auto-associative dependencies. Boruta, Decision Tree and Random Forest are used to select an object. The importance of variables is different for different methods (logistic regression, Support vector machine, Naive Bayes, XGBoost, Random Forest, neural network, decision tree). This means that the relationship between the parameters is supported only for part of the data set. That is why we propose to find the dependency for separate clusters and use this dependence for classification.

2. Literature Review

Selecting the appropriated features can be a more significant task than lessening computation time or enhancing classification or predictive accuracy. For example, in medicine [1], finding the optimal set of optimal features for the classification or predictive problem can help develop a diagnostic test.

Selecting important features (for example, determining genes appropriate for a specific type of cancer) can help decipher the mechanisms underlying the welfare problem for research. A complete enumeration of features can implement the selection method, that is, having checked all possible sets, selecting those signs for which the error is minimal. This method is simple to implement, but it is entirely ineffective on big data. Therefore, in this case, other algorithms are most often used.

The three primary classes of feature extract algorithms – filters, wrappers, and built-in algorithms are used [2].

Filters are based on some metrics that are independent of the classification method. For example, the correlation of features with the target vector and information content criteria. They are applied before classification. The most significant benefit of filtering is that it can be used as preprocessing to reduce space dimensionality and overcome overfitting. Filtering methods are generally fast. Filters are used to choose features in clustering or to build an initial approximation [3]. Unfortunately, such methods are not designed to detect complex relationships between elements and, as a rule, are not sensitive enough to specify all dependences in the data.

Embedded algorithms organize feature extraction during the classifier training approach, and it is they explicitly optimize the set of features used to achieve better accuracy [4]. The benefits of built-in algorithms are that, as a rule, they discover solutions quickly, decrease the possibility of retraining data, eliminate the need to separate data into training and test subsamples. Nevertheless, these algorithms are not ubiquitous.

Wrappers rely on feature significance information from several classifications or regression models and can thus find deeper patterns in dataset than filters. Wrappers can be built on any classifier that defines the degree of significance of the features [5].

Imputation is a procedure for assessing unknown or missing values based on available data, which allows you to form a complete set of data with some plausible estimates.

Methods of imputation are divided into non-model and model-based approaches. There are approaches based on single and multiple imputation methods in terms of the quantity of values received from imputation methods [6]. One-time filling algorithms provide a single complete set of data, where a deal replaces each space. The benefits of this method are the use of methods of comprehensive data analysis in the subsequent stages of processing. Replenishment algorithms form several complete data sets analyzed separately and later combined according to specific rules. This minimizes standard errors in the following steps by processing comprehensive datasets. Nevertheless, mentioned methods require multiple resources to create more data sets, spend more time performing the analysis, and have more memory to store the results [7].

The method based on mean substitution solves the problem of incomplete data by replacing each missing variable with an average value. There are the following types of substitutions: median value, mean value for subgroup [8], value with the highest frequency, and replacement with minimum / maximum value. This method can lead to undesirable results [8], such as variance change, negative correlation shift, and misrepresentation of the population.

Hot-deck imputation approach replaces each gap with a random value taken from an existing dataset [9]. Its significant disadvantage is the warping of correlations and covariates.

Cold-deck imputation (CD) approach implements the replacement of each gap by some constant value from an external source [9]. The specific case of CD is zero replacement. It has the same disadvantages as hot-deck imputation.

The regression model applies to replace data gaps with expected values emanated from a regression equation constructed from a complete dataset. Disadvantages of regression completion include the need to accurately define regression models, exaggerate correlation and covariance, the likelihood of moving to predicted values outside the logical sequence, and the need for extensive quantities of data to obtain consistent estimates.

The association rules (AR) mining method uses the constructed associative rules for the data imputation [10, 11]. However, the temporal complexity of this method needs to be improved [12].

Diabetes increases the likelihood of severe COVID19. New clinical data and investigations show that this might work in the opposite direction: scientists are recording new cases where COVID-19 has sharply provoked type 1 diabetes in humans. The World Health Organization views diabetes as one of the existing diseases, on a par with respectable age, making someone more vulnerable to severe COVID 19 infection. Cellular immunity is essential for protecting against viral diseases, and its effectiveness decreases with age. In particular, a decrease in T-cell receptors explains the significant increase in mortality of COVID19 with age.

Therefore, another important factor is the search for possible relationships between biomarkers of aging and COVID resistance.

Existing research and the data sets collected for this use standard machine learning methods, while demonstrating not very high accuracy of prediction. Thus, in the paper [13], there are used feature selection, XGBoost and decision tree to determine COVID biomarkers, F1-score does not rise above 0.7. In the paper [14], markers of CH4 and CH8 immunodeficiency and their association with coronavirus infections were analyzed using statistical models, including the Cox model. Accordingly, it is impossible to prevent negative situations. In paper [15], there is used the empirical mode decomposition (EEMD) and the artificial neural network (ANN) to predict the COVID-19 epidemic. Thus, in order to prolong the active period of life, it is necessary to track the dynamics of changes in molecular and biochemical markers, anthropometric indicators, behavioral factors, environmental parameters and habitat, and so on. As a result, it is necessary to use a big data-based approach to collect information from disparate datasets, process them, and further analyze them. At the same time, it is necessary to analyze small data samples, which will include time multimodal series of changes in human parameters.

The analysis of literature sources showed the lack of a comprehensive approach to solving the problem of prolonging the active period of life and preventing exacerbation of chronic diseases. At the same time, as we see in the case of COVID, chronic diseases (diabetes and obesity) can not only reduce the ability to work, but also increase the likelihood of severe course of other diseases.

3. Materials and Methods

Biomarkers of aging are used to predict possible changes in the body that lead to disability due to functional age-related changes. Biomarkers of aging are markers that can predict the functional capacity of an organism at a certain age better than chronological age. Since the immune system is a change the language here for English version leading factor in aging, the main impact of which is realized through increased inflammation and reduced effectiveness of cellular immunity, it becomes clear the need to involve relevant markers to develop interventions to increase the duration of healthy longevity.

Thus, control of chronic age-related diseases (diabetes and obesity) and biological markers can be used to predict functional changes in the body, and analysis of other personal indicators will determine how to reduce the negative effects of such changes by extending the period of active longevity.

Sociological research also shows that people in certain regions remain active for a long time and there are far fewer people who are obese. Therefore, it is also advisable to analyze the parameters of the environment and habitat and its impact on the parameters of the organism.

The baseline of the hybrid ensemble feature selection model looks like the following:

- Several selectors using,
- Aggregation of the results.

Several wrapper algorithms will be used in the preprocessing stage for the first stage.

The correlation matrix shows the numerical value of the correlation coefficient for all possible combinations of variables. It is mainly used to find out the relationship between more than two variables. The decision tree returns the feature weight as the criterion for evaluating features. It allows building a ranked list of selected features using different measures. Classification and Regression Trees (CART) was used for feature selection with Gini-index as a measure in our case.

Random Forest is an ensemble of numerous training-sensitive algorithms (decision trees). Mentioned approach has a slight compensation. The bias of the training method is the deviation of the average response of the trained algorithm from the reaction of the ideal algorithm. Each of these classifiers is built on a random subset of entities and a random subset of characters.

Boruta is a heuristic algorithm for choosing important features based on Random Forest approach [16]. Features with the Z-measure less than the maximum Z-measure among the added features are removed at each iteration. To calculate the Z-measure, we need to calculate its importance, obtained using the built-in algorithm in Random Forest, and divide it by the standard deviation of the feature importance. Added features are obtained as follows: the characteristics available in the selection are copied, and then every new attribute is filled by mixing its values. This procedure is repeated several times to get statistically significant results, and variables are generated independently at each iteration.

The Jaccard index [17] measures the similarity of the feature subsets chosen by separated feature selectors (each selector is organized as a separated iteration):

$$(S_1, \dots, S_{1n}) = \frac{|S_1 \cap \dots \cap S_n|}{|S_1 \cup \dots \cup S_n|}, \tag{1}$$

where S_i is the subset of features at the i-th iteration, for i=1,..,n. The value of the Jaccard index varies from 0 to 1, where 1 implies the absolute similarity of subsets.

The schema of the hybrid ensemble feature selection model is given in Fig. 1.

Next, the method of missing data imputation is developed. This method is based on classical functional dependencies in relational databases and association rules from non-relational databases. It consists of two parts:

- Probabilistic Production Dependencies mining;
- the Probabilistic Production Dependencies usage for missing data imputation.

Investigating extensive data requires specifying attribute clusters that form functional dependencies (FD) [18]. However, datasets are extensively standard in the real world, with essential dependencies defined only on a subset of crucial attribute group values. Moreover, the reliance can appear not only for tuples in relational data sources but also between subsets of values in different tuples. We will name them Probabilistic Production Dependency (PPD).

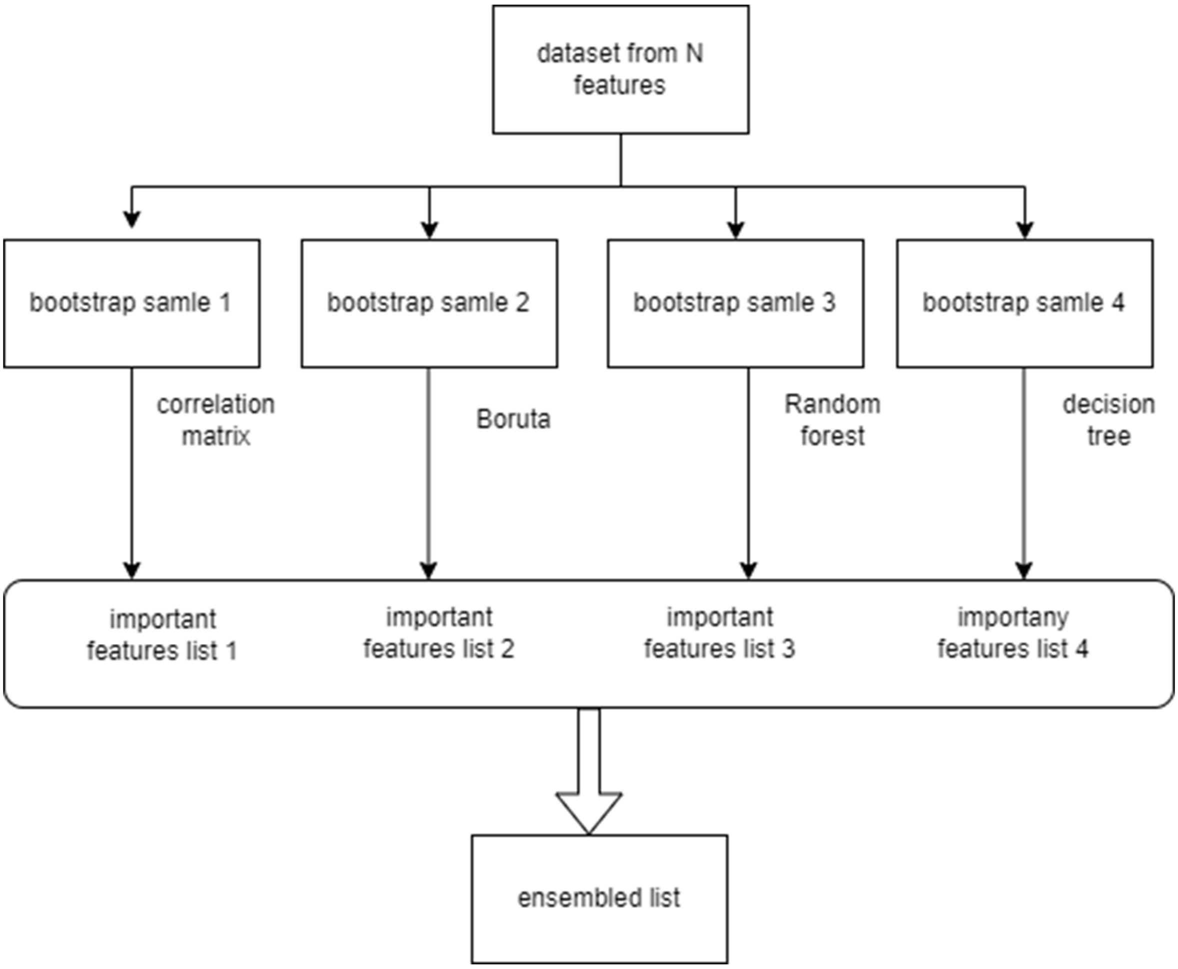

Figure 1: The hybrid ensemble feature selection model.

Probabilistic Production Dependency is a dependence similar to associative rule in the primary ratio selection that is proper for many entities. The significance threshold should be specified expertly or based on estimation of the probability of erroneous selection of this dependence. The main distinction between associative rules and PPD is that PPD will be generated from existing functional dependencies (FD) in the dataset [9].

$$\begin{gathered} F_I : \ K = \{a_i\}, a_i \in A, D = \{a_j\}, \\ a_j \in A,: P(k \in K \to d \in D) = p \end{gathered}, \tag{2}$$

where k and d are the tuples of groups of attributes K and D, respectively.

The gaps and missing data presented among the values of the attribute Y of the relation r are classified using PPD. The following algorithm for PPD mining is proposed.

```
Algorithm 1. PPD mining algorithm
 1. Entities with the same X-values will be grouped;
 2. To choose attributes from FD with same X and add them to Y;
 3. To calculate the Support and Confidence, Imputation of the
    tuple selected in step 2);
 4. To identify the tuples with the highest value of Confidence;
 5. To add X→Y to PPDset.
```

To fill in missing data, the PPD should be built.

Next, the novel algorithm for missing data imputation is developed.

```
Algorithm 2. Data imputation algorithm
Completeness=0
While Completeness/100< Imputation
Arrange all attributes from PPDset by Confidence level
For each group:
  If percentage of non-empty Y-value is higher or equal to Support
    fill in empty values using PPDset
    Completeness++
  Else
    Merge PPD using Armstrong rules
```

Next, a hierarchical classifier as a two-stage data prediction algorithm is developed. The first stage is clustering; the next step is to build a classification model for each separate cluster. K-means [19] together with the random forest do not dominate other cluster models. The hierarchical classifier is constructed as follows:

- the appropriate number of clusters was found using gap statistics;
- the density of distribution is calculated;
- XGboost and Random Forest are used for each cluster separately;
- hard voting for the results. Based on it, the class with the highest number of votes will be selected. If the voices are the same, the result of the classifier with the minimum value of depth will be selected.

4. Experiments and results

To validate the proposed methods, dataset with medical data is used.

Dataset consists of 35 features and 122 instances collected from Lviv regional rehabilitation center for post-COVID patients with short- and long-term (more than 20 days) treatment and rehabilitation.

The personal data were removed from the dataset and replaced with unique random identifiers. The next feature, sex, is processed using one-hot encoding technics and in the final dataset is given in two components – female and male. Features like age, weight, height, BMI, CAT, pulse, the function of external respiration are taken as physiological parameters measured before inpatient treatment. The rest of the features were immune-based biomarkers as described below.

Zero cells (0-lymphocytes) do not carry T- and B-cells markers. Zero cells make 10–20 % of the total lymphocytes in human peripheral blood. Some researchers consider them immature or overripe T- or B-lymphocytes because they have a small number of antigens common to B- and T-cells. Zero cells include K-cells and NK-cells.

CD3+ is a surface marker specific to all T-lymphocyte subpopulation cells. By function, it belongs to the family of proteins that form a complex of membrane signaling associated with the T-cell receptor. Mature T-lymphocytes are "responsible" for cellular immune reactions and conduct immunological monitoring of antigenic homeostasis in the body.

CD4+ is a characteristic of helper T-cells; also represented on monocytes, macrophages, dendritic cells. It binds to class II MHC molecules expressed on antigen-presenting cells, facilitating the recognition of peptide antigens. Helper T-lymphocytes (CD4+) are helpers (inducers) of the immune response, cells that regulate the strength of the body's immune response to a foreign antigen, as well as control the stability of the body's internal environment (antigenic homeostasis) and cause increased antibody synthesis [20].

CD8+ is a characteristic of suppressor and cytotoxic T-cells, NK-cells, mostly thymocytes. It is a T-cell activation receptor that facilitates the recognition of cell-bound class I MHC antigens.

CD16+ natural killers are part of innate immunity; they are involved in early response against viral infections and intracellular bacteria. Compared with cells of specific immunity (T- and B-lymphocytes), they have the advantage that they do not require long-term activation. Besides, NK-cells complement the action of T-cytotoxic cells can also regulate the immune response by producing various cytokines, including interferon-γ. They are the primary cells of antitumor protection. Their role is vital in manifesting cellular immunity in viral, protozoan, fungal and bacterial diseases caused by intracellular parasites. Their action is enhanced by interferon. The functions performed by natural killers can be divided into two main types: the production of cytokines that regulate the work of other cells of the immune system and the direct destruction of damaged cells.

Mature B lymphocytes express CD22+ markers. B-lymphocytes are responsible for the humoral adaptive immune response, primarily at removing extracellular infectious agents. After binding to a specific antigen, B-lymphocytes, in cooperation with T-lymphocytes and T-helpers proliferate, differentiate into plasma cells that secrete antibodies/immunoglobulins and memory cells. Defects of humoral immunity associated with the B-cells are sporadic, so a common hypoimmunoglobulinemia is mainly caused by other reasons.

CD4/CD8 immunoregulatory index reflects the ratio of CD4+ cells (T-helpers) to CD8+ cells (T-cytotoxic cells). It is a relative indicator that has an indicative value. Its small increase or decrease has no independent diagnostic value. Changes in the index significance the clinician to focus on the reasons for the deviation of this index. The immunoregulatory index is assessed relative to the phase of the immune response. In the period of exacerbation and remission of clinical manifestations, the immunoregulatory index reaches high values due to the high percentage of T-helpers (CD4+ T-cells). During the recovery period, the indicator's value decreases due to the increase in the level of CD8+ T-cells (killers). Violation of this pattern indicates the inadequacy of the immune response and the possibility of chronic infection due to incomplete removal of the pathogen [21].

We implemented our approach in Rstudio. The essential packages we used were caret, rpart, Metrics, Boruta, Random forest, rules and ggplot2 for visualization. To generate PPD, the minimal support threshold equal to 0.0001 in the a priori algorithm is chosen. In addition, all rules with confidence below 0.001 are filtered out.

First, the predictive accuracy for whole dataset and selected features was analyzed (Table 1).

Table 1
The comparison of ML models prediction results using the whole set of features and selected features subset

Model	For the whole dataset			For selected features		
	MSE	MAE	R2	MSE	MAE	R2
k-NN	7.029	2.155	-0.162	5.781	1,825	0.044
SVM	5.753	1.828	0.049	5.521	1.812	0.038
SGD	16.007	3.280	-1.647	12.328	2.603	-1.039
Linear Regression	17.826	3.571	-1.948	14.836	2.760	-1.453
MLP NN	5.717	1.777	0.055	5.469	1.034	0.034

The much higher accuracy is obtained with SVM (Support vector machine, polynomial kernel) and artificial neural network (one hidden layer with 12 neurons in it, sigmoid activation function) for selected features [22]. The used measures are the following [23]:

- mean squared error MSE,
- mean absolute error MAE,
- R2.

Next, missing data imputation method is used. The developed method was compared with the existing ones: associative rules (AR), random forest (RF), support vector machine (SVM), multilayered perceptron (MLP), expectation-maximization (EM) and k-nearest neighbor (KNN) (Fig. 2). The recovery error is presented using normalized root-mean-square error (NRMSE).

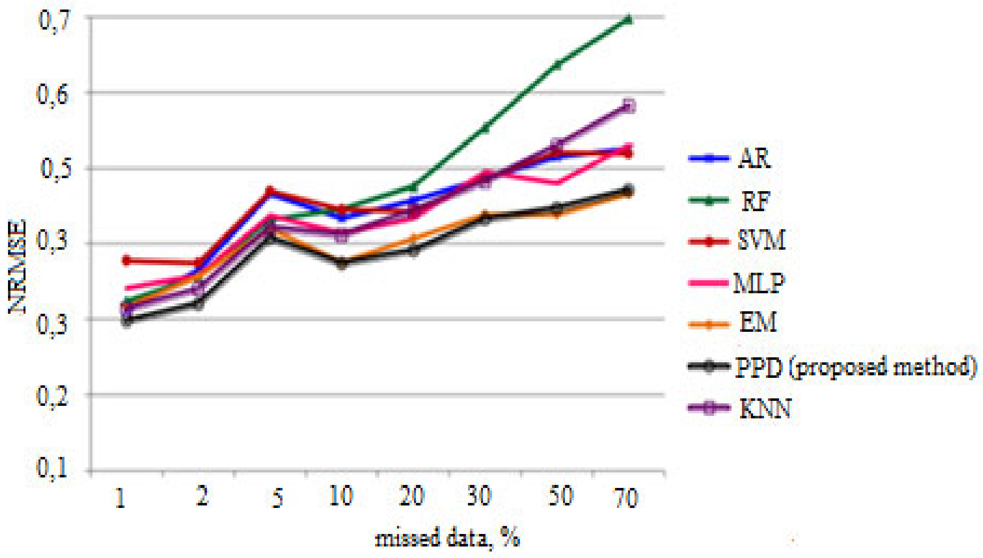

Figure 2: NRMSE recovery error.

The proposed missing data imputation method creates additional data values operating a based domain and functional dependencies and includes these values to available training data. The correctness of the filled-in values is proved on the predictor built on the original dataset. The proposed PPD method conducts 12% better than RF and EM models for 30% missing data.

5. Conclusion

Current trends in the development of information technology and databases are analyzed. As a result, unsolved problems in the field of dependency search in large databases of several subject areas, in particular, in medicine, were revealed. The analysis of existing methods and means of detection of dependences in data is carried out. This made it possible to identify a new subclass of dependencies - Probabilistic Production dependencies. A method for deriving PPD in relational databases has been developed.

A new hybrid ensemble feature selection model for a machine learning-based post-COVID prediction system is proposed as an automatic feature cut-off rank identifier. Ensemble methods for processing of multimodal data based on a hierarchical classifier, a set of weak classifiers and a number of methods for selecting important characteristics with a much higher value of accuracy on unbalanced data sets compared to existing machine learning methods are developed.

The associative rules are found together with weak predictors usage to improve the classification quality.

The proposed missing data imputation method creates additional data values operating a based domain and functional dependencies and includes these values to available training data.

The correctness of the filled-in values is proved on the predictor built on the original dataset. The proposed PPD method conducts 12% better than RF and EM models for 30% missing data. The EM method looks the best for more additional missing data (the range of about 40%-50% missing data), and the PPD has equivalent results with the SVM (support vector machine).

Comparison of approaches for investigating and modeling the statistical processes by qualitative criteria verified the proposed method has the subsequent benefits:

- retaining the characteristics of resistance to errors in the data;
- allowing the parallel implementation in distributed databases;
- automating and performing the analysis of various data types.

6. References

[1] M.B. Kursa, W.R. Rudnicki, The all relevant feature selection using random forest, arXiv preprint arXiv:1106.5112, 2011.

[2] G. Chandrashekar, F. Sahin, A survey on feature selection methods. Computers Electrical Engineering, 40(1), 16-28 (2014).

[3] A. Bommert, X. Sun, B. Bischl, J. Rahnenführer, M. Lang, Benchmark for filter methods for feature selection in high-dimensional classification data. Computational Statistics Data Analysis, 143, 106839 (2020).

[4] B. Venkatesh, J. Anuradha, A review of feature selection and its methods. Cybernetics and Information Technologies, 19(1), 3-26 (2019).

[5] L. N. Sanchez-Pinto, L. R. Venable, J. Fahrenbach, M. M. Churpek, Comparison of variable selection methods for clinical predictive modeling. International journal of medical informatics, 116, 10-17 (2018).

[6] P. Hayati Rezvan, K. J. Lee, J. A. Simpson, The Rise of Multiple Imputation: A Review of the Reporting and Implementation of the Method in Medical Research. BMC Med Res Methodol, , 15, 30 (2015). https://doi.org/10.1186/s12874-015-0022-1

[7] N. Ahmat Zainuri, A. A. Jemain, N. A Muda, Comparison of Various Imputation Methods for Missing Values in Air Quality Data. JSM, 44, 449–456 (2015). https://doi.org/10.17576/jsm-2015-4403-17

[8] C. A. Leke, T. Marwala, Introduction to Missing Data Estimation. Deep Learning and Missing Data in Engineering Systems, 1-20 (2019). https://doi.org/10.1117/12.2053057

[9] C. Wang, N. Shakhovska, A. Sachenko, M. Komar, A New Approach for Missing Data Imputation in Big Data Interface. Information Technology and Control, 49(4), 541-555 (2020).

[10] M. Azmi, G. C. Runger, A. Berrado, Interpretable regularized class association rules algorithm for classification in a categorical data space. Information Sciences, 483, 313-331 (2019).

[11] F.Thabtah, P. Cowling, Y. Peng, MCAR: multi-class classification based on association rule. In The 3rd ACS/IEEE International Conference on Computer Systems and Applications, (2005)/

[12] K. Mittal, G. Aggarwal, P. Mahajan, A comparative study of association rule mining techniques and predictive mining approaches for association classification. International Journal of Advanced Research in Computer Science, 8(9) (2017).

[13] L. Yan, H. T. Zhang, J. Goncalves, Y. Xiao, M. Wang, Y. Guo, Y. Yuan, An interpretable mortality prediction model for COVID-19 patients, Nature machine intelligence, 2 (5), 283-288 (2020).

[14] A. Trickey, M. T. May, P.Schommers, J. Tate, S. M. Ingle, J. L. Guest, J. A. Sterne, CD4: CD8 ratio and CD8 count as prognostic markers for mortality in human immunodeficiency virus – infected patients on antiretroviral therapy: the Antiretroviral Therapy Cohort Collaboration (ART-CC), Clinical Infectious Diseases, 65 (6), 959-966 (2017).

[15] N. Hasan, A Methodological Approach for Predicting COVID-19 Epidemic Using EEMD-ANN Hybrid Model. Internet of Things 11, 100228 (2020). https://doi.org/10.1016/j.iot.2020.100228

[16] L. N. Sanchez-Pinto, L. R. Venable, J. Fahrenbach, M. M. Churpek, Comparison of variable selection methods for clinical predictive modeling. International journal of medical informatics, 116, 10-17 (2018).

[17] A., Dariush, A. Bastanfard, An objective method to evaluate exemplar-based inpainted images quality using Jaccard index, Multimedia Tools and Applications 80.17, 26199-26212 (2021).

[18] A., Federico, et al., Association rules extraction for the identification of functional dependencies in complex technical infrastructures, Reliability Engineering System Safety 209 (2021).

[19] I.-D. Borlea, et al., A unified form of fuzzy C-means and K-means algorithms and its partitional implementation, Knowledge-Based Systems 214 (2021): 106731.

[20] A.M. Miggelbrink, et al., CD4 T-cell exhaustion: Does it exist and what are its roles in cancer?, Clinical Cancer Research 27.21, 5742-5752 (2021).

[21] L. I. U. Huan-xia, et al., Analysis of CD4/CD8 ratio in HIV-infected patients who accepted initial antiretroviral therapy for 48 weeks, China Tropical Medicine 21.3 (2021).

[22] D.A. Otchere, et al., Application of supervised machine learning paradigms in the prediction of petroleum reservoir properties: Comparative analysis of ANN and SVM models, Journal of Petroleum Science and Engineering 200 (2021): 108182.

[23] D. Chicco, M. J. Warrens, G. Jurman, The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. PeerJ Computer Science, 7, e623 (2021).