Обґрунтовано підхід до імпутації на основі розширеної корельованої матриці атрибутів. Обґрунтовано необхідність попередньої кластеризації даних для зменшення значної кількості унікальних значень. Виконано синтез моделей імпутації на основі машинного навчання і досліджено їхню ефективність, виконано їх порівняння із методом ампутації Most Common Value. Розроблено ансамблі моделей імпутації для числових і номінальних даних. Показано, що ансамблі дають найбільш стійкі й ефективні результати ампутації

Ключові слова: імпутація даних, розширена матриця атрибутів, модель імпутації, ансамбль моделей

Обоснован подход к импутации на основе расширенной коррелированной матрицы атрибутов. Обоснована необходимость предварительной кластеризации данных для уменьшения значительного количества уникальных значений. Выполнен синтез моделей импутации на основе машинного обучения и исследована их эффективность, выполнено их сравнение с методом импутации Most Common Value. Разработаны ансамбли моделей импутации для числовых и номинальных данных. Показано, что ансамбли дают наиболее устойчивые и эффективные результаты импутации

Ключевые слова: импутация данных, расширенная матрица атрибутов, модель импутации, ансамбль моделей

UDC 303.7:004.6
DOI: 10.15587/1729-4061.2016.74871

DEVELOPMENT OF MODELS FOR IMPUTATION OF DATA FROM SOCIAL NETWORKS ON THE BASIS OF AN EXTENDED MATRIX OF ATTRIBUTES

O. Slabchenko*
E-mail: slabchenko.olesia@gmail.com
V. Sydorenko
PhD, Associate Professor*
E-mail: vnsidorenko@gmail.com
X. Siebert
Associate professor, PhD
Mathematics and Operational Research Department
University of Mons
rue de Houdain, 9, Mons, Belgium, 7000
E-mail: xavier.siebert@umons.ac.be
*Computer and information systems department
Kremenchuk Mykhailo Ostohradskyi National University
Pershotravneva str., 20, Kremenchuk, Ukraine, 39600

1. Introduction

In recent years, dramatic progress and popularity of the Internet and social networks in particular have caused the accumulation of a substantial volume of data available for analysis. The ability to collect such data and the availability of technological means for this purpose have resulted in a meaningful shift in social network analysis (SNA) and data mining (DM) techniques being employed to improve business processes and to develop special services for users. On the basis of data from social networks, the following tasks are solved: building of recommender systems, organization of mechanisms for interaction with customers, advertising and promotion of products and services, searching and recruiting of experts, trend monitoring, etc. However, studies in this area show that one of the most common problems in all types of social network analysis still remains the poor quality of data, which complicates their analysis [1, 2]. The main factor influencing the poor quality of raw data is the essential number of missing values caused by a specific character of data acquisition and storage in social networks (any field may contain gaps, for instance, "Age/Date of birth", "Sex", "Marital status"). A solution to the problem of data quality and completeness is a very important preparatory stage, and, according to the CRISP-DM methodology, it should precede the stage of modelling. This is because the DM and SNA techniques and algorithms are unable to handle datasets with missing values: they require data that meet the conditions of completeness. Incorrect treatment of missing data, including ignoring or deleting, may cause a range of negative consequences [3]. Therefore, correct treatment of initial data is an urgent problem at the stage of pre-processing because adequacy and reliability of further modelling results depend on it.

2. Analysis of scientific literature and the problem statement

The choice of a method for incomplete data treatment depends on the missing data mechanism. In previously published studies, the mechanisms of missing data are traditionally divided into three classes [4, 5]: missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). In case the missing data mechanism is beyond a researcher's control, it is recommended to assume a MAR mechanism [6]. This is due to the fact that an erroneous assumption of an MCAR mech-

anism when using deletion [7] or weighting methods [8] for incomplete data leads to obtaining non-representative samples and reducing statistical power, substantial errors and biased results of analysis in presence of statistically significant correlations between variables [3]. As opposed to the methods mentioned above, which have nothing to do with filling in gaps, methods of imputation produce missing values estimation [9]. Herewith, imputation methods utilize all available data and allow obtaining a complete dataset that is suitable for further analysis. Imputation methods can correctly deal with MCAR and MAR mechanisms [10]; hence, they are suitable for a wide range of application areas.

The group of imputation methods includes a large number of algorithms for estimation of missing values: from simple approaches based on substitution of certain values [11] to the use of model methods [12, 13] that approximate dependence of missing values on the obtained data. The most promising are methods of imputation on the basis of data mining algorithms, which are able to discover internal patterns in data and use them for the further process of imputation [14, 15]. However, analysis of state-of-the-art studies and modern trends in the imputation area shows that it is extremely difficult to develop a universal model that could show good performance in various subject areas. Therefore, in many research papers, models are mainly offered for specified domains [16–18]. They include specific algorithms for data treatment on the basis of machine learning techniques and take into account the specific nature of the presented data domain [19, 20].

Previously published studies show that attributes from social networks are mixed-type and contain a large number of unique values. Application of the well-known methods of imputation – on the basis of both simple substitution and classical model methods is complicated by the factors mentioned above [21], and it does not always produce the desired result [1]. Moreover, such methods are not considered in the context of complex networks, which are social networks, and they do not recognize real connections between accounts and the specifics of data organization in social networks.

Therefore, an important issue is to develop specialized methods for imputation of data from social networks on the basis of approaches that are able to solve the problem of a large number of unique values in attributes and to utilize all available significant data from users' accounts.

3. The goal and objectives of the study

The goal of this work is to improve the quality of imputation of missing data from social networks' accounts through the design of models and ensembles of models for imputation while using an extended matrix of attributes.

To achieve this goal, the following tasks should be solved:

– to research the main features of data from social networks' accounts and to justify an approach to imputation on the basis of an extended matrix of attributes;

– to develop models of imputation on the basis of data mining algorithms with introduction of a pre-clustering step;

– to develop ensembles of models of imputation for handling data of various nature and to research their performance.

4. Materials and methods of imputation of incomplete data from social networks' accounts

4. 1. Features of real data from social networks' accounts

Research shows that users' profiles usually contain, besides standard personal details, other specific data such as text and multimedia content, activity logs, data about profile settings, and relationship links, which are also available for analysis. Presence of such data determines the main positive features of indicators in users' accounts, such as minuteness (a wide range of indicators describing users in detail) and retrospection (availability of biographical data) [22]. Nevertheless, a set of mixed-type indicators, on the one hand, and specific characteristics of data acquisition and storage, on the other hand, cause a range of substantive problems hampering their analysis. These problems include:

– high-dimensional mixed-type data (numerical, dichotomous, categorical and multimedia information);

– ill-structured or unstructured indicators as a result of lack of a unified data format;

– anomalies, inconsistency, incorrect and missing values;

– a large number of unique values in categorical indicators and a wide sample span in numerical ones.

Analysis of the structure of data from social networks' accounts shows that the problem of a high rate of missing values may arise at any stage of social networks' analysis (data collection, detection of causal dependence based on the available data, or generalization of the obtained results) [1]. The type of data incompleteness depends on the character of the problem being solved; it is conditioned by the settings of the data search filter for a particular sample of data being collected. In general, the percentage of incomplete data can exceed 80 % and have an adverse effect on the analysis. However, not all indicators from users' accounts may be incomplete. Generally, they can be divided into two groups: attributes that may potentially contain missing values and the ones that are always complete according to their nature and definition. Potentially incomplete data include a range of indicators specified personally by a user during registration or while using a social network; always complete data are attributes that are generated, stored and used by a social network system regardless of users but are also available for collection and analysis. Information about profile settings, activity and multimedia content belongs to always complete data. Let us introduce the concept of a matrix $D=k'n$ describing n users in a space of k features and referring to D as an extended matrix. It can be defined as a unity of two subsets: an incomplete matrix $X_1=k_1'n$, where k_1 is the number of indicators, which may contain missing values, and an enrichment matrix $X_2=k_2'n$, where k_2 is the number of always complete attributes: $D=X_1+X_2$, where $k=k_1+k_2$. The problem of imputation can be stated as follows: we need to obtain estimates of missing values of a matrix X [21] on the basis of the incomplete matrix X_1 and the enrichment matrix X_2.

Let us consider VK, which is the most popular social network in post-Soviet countries, to research the structure and completeness of indicators from real social networking profiles. To do this, we will compile a sample of data based on information from users' accounts that, for instance, belong to Lviv Polytechnic University. Thus, we will get an initial data set $D=12'13198$ with a comprehensive view of different indicators: age, gender, social status, geographical dispersion, etc.

V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12
NA	NA	1980	1974	4	0	0	1	4	0	0	1
NA	NA	2012	NA	7	0	0	1	4	0	0	1
29	7	2009	NA	11	0	202	1	2	62	0	1
27	2	NA	NA	78	0	139	1	4	196	0	1
NA	NA	2001	1995	4	0	11	0	4	6	0	1
34	NA	2003	1998	49	3	230	1	4	1,370	1	1
24	NA	NA	1998	13	0	0	0	3	21	0	1
...	...	...	...	...	...	...	...	...	...	...	...
The incomplete matrix X_1				The enrichment matrix X_2							

Fig. 1. A fragment of a real data sample from social networking profiles

Fig.1 illustrates a fragment of an obtained data sample where attributes from users' accounts are designated as V1 through V12 and the missing values are indicated as NA (not accessible). As we can see, the incomplete matrix X_1 consists of $k_1=4$ attributes V1 through V4 with omissions, and the enrichment matrix X_2 includes $k_2=8$ complete attributes V5 through V12.

4. 2. Research of the structure and correlations between the attributes

Having such a subset of absolutely complete data, it is reasonable to test the hypothesis about a relationship between the indicators of the matrices X_1 and X_2 in order to include the enrichment matrix X_2 in further analysis. To research dependencies between them, we use a correlation analysis, including construction of a correlation matrix based on the data sample and test the hypothesis concerning statistical significance of the estimated correlations. The choice of methods to estimate the correlation depends on the type of indicators being analysed (Table 1).

Table 1

The types of scales of available attributes

Attribute code	Attribute name	Type of scale
V1	Age	Ratio
V2	Marital status	Ordinal
V3	Year of graduation from university (UG)	Interval
V4	Year of leaving school (SG)	Interval
V5	Number of followers	Ratio
V6	Number of subscriptions	Ratio
V7	Number of friends	Ratio
V8	Sex	Dichotomous
V9	Openness	Interval
V10	File activity	Interval
V11	Communicativeness	Interval
V12	Index of website attendance	Interval

To assess a linear relationship between the variables with an interval or ratio scale, we use the Pearson product-moment correlation coefficient ρ_p. Its point estimate is defined as:

$$r_\rho = \frac{\sum_{i=1}^{n}\left(x_i-\bar{x}\right)\left(y_i-\bar{y}\right)}{\sqrt{\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2\sum_{i=1}^{n}\left(y_i-\bar{y}\right)^2}}, \quad (1)$$

where $\underline{x_i}$ and $\underline{y_i}$ are series of n measurements of X and Y; $\bar{x}$ and $\bar{y}$ are sample means of X and Y.

Hence, to research the relationships between all $k=k_1+k_2$ attributes of the matrix D (except for "Marital status" and "Sex"), we assess the point estimate of the linear Pearson correlation r_p (Table 2).

To determine whether the obtained correlations are statistically significant, we use a significance level of $\alpha=0.05$. All correlations in Table 2 are statistically significant, except for the ones highlighted in colour. It can be seen from the estimated correlations that the extended matrix D is characterized by the following correlations: a weak but statistically significant correlation between the groups of complete and incomplete attributes from X_1 and X_2; a strong correlation between potentially incomplete attributes within X_1; correlations between always complete attributes within X_2.

To research the dependencies between the attribute "Marital status" and the other indicators, we use the Spearman's rank correlation coefficient ρ_s between the ranked variables, which is usually applied for ordinal scales. Its point estimate is defined as follows:

$$r_s = 1-\frac{6\sum(D)^2}{n\left(n^2-1\right)}, \quad (2)$$

where n is the number of indicators that are analysed, D is the difference between the two ranks of each indicator, and $\sum(D)^2$ is the sum of squared differences between the ranks.

As we can see from Table 3, "Marital status" is characterized by a weak but statistically significant correlation with the attributes from the incomplete matrix X_1 and the attribute "Communicativeness" from the enrichment matrix X_2.

Thus, the research of the structure and correlations between the $k_1=4$ attributes with omissions and the $k_2=8$ complete attributes shows a statistically significant relationship between the incomplete matrix X_1 and the enrichment matrix X_2. The existence of this relationship confirms that the use of the extended matrix D is suitable for further development of improved methods for imputation of missing data.

Table 2

A matrix of Pearson correlations between the complete and incomplete attributes

The variables	The subset of incomplete data of X_1			The subset of complete data (enrichment matrix) of X_2						
	V1	V3	V4	V5	V6	V7	V9	V10	V11	V12
V1	1	–0.845	–0.805	–0.035	–0.025	–0.171	0.172	–0.215	–0.203	0.048
V3	–0.845	1	0.833	0.036	0.034	0.177	–0.169	0.229	0.197	–0.057
V4	–0.805	0.833	1	0.036	0.040	0.176	–0.177	0.220	0.195	–0.064
V5	–0.035	0.036	0.036	1	0.025	0.118	–0.032	0.122	0.040	–0.008
V6	–0.025	0.034	0.040	0.025	1	0.349	–0.008	0.093	0.036	0.001
V7	–0.171	0.177	0.176	0.118	0.349	1	–0.139	0.415	0.171	–0.024
V9	0.172	–0.169	–0.177	–0.032	–0.008	–0.139	1	–0.152	–0.093	0.039
V10	–0.215	0.229	0.220	0.122	0.093	0.415	–0.152	1	0.256	–0.047
V11	–0.203	0.197	0.195	0.040	0.036	0.171	–0.093	0.256	1	–0.047
V12	0.048	–0.057	–0.064	–0.008	0.001	–0.024	0.039	–0.047	–0.047	1

Table 3

A matrix of Spearman correlations between the complete and incomplete attributes

The variables	The subset of incomplete data of X_1				The subset of complete data (enrichment matrix) of X_2						
	V1	V2	V3	V4	V5	V6	V7	V9	V10	V11	V12
V1	1	0.310	–0.917	–0.879	–0.517	–0.266	–0.464	0.235	–0.419	–0.297	0.091
V2	0.310	1	–0.292	–0.277	–0.092	–0.078	–0.039	0.090	–0.080	–0.120	0.050
V3	–0.917	–0.292	1	0.858	0.503	0.263	0.460	–0.224	0.418	0.295	–0.093
V4	–0.879	–0.277	0.858	1	0.491	0.262	0.432	–0.216	0.394	0.260	–0.090
V5	–0.517	–0.092	0.503	0.491	1	0.352	0.693	–0.258	0.674	0.260	–0.081
V6	–0.266	–0.078	0.263	0.262	0.352	1	0.532	–0.147	0.474	0.184	–0.041
V7	–0.464	–0.039	0.460	0.432	0.693	0.532	1	–0.299	0.844	0.277	–0.038
V9	0.235	0.090	–0.224	–0.216	–0.258	–0.147	–0.299	1	–0.285	–0.142	0.030
V10	–0.419	–0.080	0.418	0.394	0.674	0.474	0.844	–0.285	1	0.310	–0.049
V11	–0.297	–0.120	0.295	0.260	0.260	0.184	0.277	–0.142	0.310	1	–0.051
V12	0.091	0.050	–0.093	–0.090	–0.081	–0.041	–0.038	0.030	–0.049	–0.051	1

4. 3. Models of imputation building

Besides the missing values, real data from social networks may also contain a large number of unique values in attributes, which complicates analysis. Taking into account the revealed relationships between the attributes, it is reasonable to assume that users can form some groups based either on similar personal details or activity patterns and thus generate natural clusters. So it is reasonable to use clustering techniques when developing models based on such data: either for reconnaissance studies of initial data in order to find potential clusters of objects or for a dimensionality reduction of analysable data.

It is obvious that k=12 input indicators are difficult to analyse; they need pre-processing. To reveal latent generalizing characteristics of the structure of the studied data and to reduce their dimensionality, we use an approach on the basis of factor analysis with a further step of clustering [1]. Thus, we obtain three orthogonal factors accounting for 94.87 % of the variability in the original k=12 variables. The obtained loadings can be defined as follows: the 1st factor is the age profile, the 2nd is the account strength, and the 3rd is communicativeness. The next step of the clustering allows deriving clusters with the following structure: the 1st cluster consists of 529 observations, the 2nd one consists of 1,139 observations, the 3rd one consists of 434 observations, the 4th one consists of 520observations, and the 5th one consists of 603 observations. An observation inside the clusters proves that objects with similar values of indicators are grouped. We have found that the number of unique values in the attributes decreases considerably inside the clusters in comparison with the unclustered data. Table 5 and Fig. 2 show information about the number of unique values in the attributes of the largest 2nd cluster and the unclustered data.

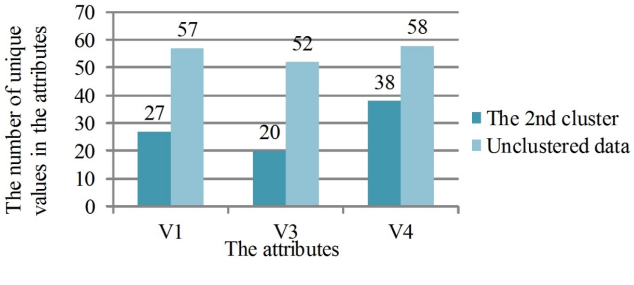


Fig. 2. The number of unique values in the attributes V1, V3, and V4 after clustering

Table 4

The number of unique values in attributes

The attributes	The number of unique values	
	The 2nd cluster (1,139 observations)	The unclustered data D
V1	27	69
V2	7	8
V3	20	58
V4	38	64
V5	114	325
V6	36	100
V7	326	640
V8	2	2
V9	5	5
V10	391	979
V11	4	6
V12	6	7

The diagram shows that the use of the suggested approach makes it possible to decrease the number of unique values in potentially incomplete attributes inside the obtained clusters by 40–65 %. Thus, the existence of users' groups with similar values of indicators helps use clustering in order to reduce the dimension of the variables under consideration. The proposed approach to the reduction of the dimension and the decrease in the number of unique values can be represented as follows (Fig. 3).

Algorithm 1. Reduction of the number of unique values in the attributes on the basis of clustering:

Input: $k \times n$ of the matrix D.

1: Reduction of the dimensionality of the extended matrix D and a transformation into D' of a smaller dimension;
2: Clustering of D';
3: $s \leftarrow$ the number of clusters;
4: ***for*** q *in* s ***do***:
5: Construction of the subset of data from the current cluster D'_q;
6: Restoration of the dimensionality of the matrix D'_q and a transformation into D_q of the initial dimension;
7: ***return*** s subsets of D_q data from q clusters.

Fig. 3. An algorithm of reducing the number of unique values in the attributes

The use of the proposed approach makes it possible to get a simpler structure of the indicators describing users of social networks inside the derived clusters and facilitates

further processing and analysis of data, including imputation of missing values.

Studies show that data from users' accounts are mixed-type. It determines the choice of the methods to process them: these methods must be able to deal with different types of variables simultaneously; otherwise, we have to exclude indicators that cannot be processed by specified methods. For imputation of data from accounts of social network users, we have developed the following models, including the suggested approach of data pre-processing on the basis of pre-clustering.

1. A model on the basis of association rules (AR): it is built on our improved algorithm [1]. The main parameters determining its work are minimum support s_{min} and confidence c_{min} of the generated rules. The algorithm of the AR model looks as follows (Fig. 4).

Algorithm 2. Imputation of missing values on the basis of association rules:
Input: $k \times n$ of the matrix D, the minimum support s_{min}, and the minimum confidence c_{min}.
1: Clustering of D;
2: $s \leftarrow$ the number of clusters;
3: ***for q*** *in* ***s do***:
4: Replacement of missing values in D_q with the special value "missing" and obtaining D_q^{obs};
5: Search of frequent itemsets L_q in D_q^{obs};
6: Generation of candidates C_q from transactions $t \in T$;
7: Selection of candidates L_q with the support $s_q \geq s_{min}$;
8: Generation of association rules R_q from candidates L_q with the confidence $c_q \geq c_{min}$;
9: $m = k_1 \leftarrow$ the number of attributes with omissions;
$l \leftarrow$ the number of missing values;
10: ***for i*** *in* ***m do***:
11: Selection of rules R_q^i from R_q that contain the current attribute i in the consequence $cons(R_q^i)$;
12: Sorting of R_q^i by the confidence c_q in a descending order;
13: ***for j*** *in* ***l do***:
14: Search of a suitable rule for imputation of an ***j***-th omission with the antecedent $antec(R_q^i)$;
15: Imputation of the missing value d_{ij} by the respective value from the consequence $cons(R_q^i)$;
16: ***end for***
17: ***end for***
18: ***end for***
19: ***return*** matrix with imputed values D^{imp}.

Fig. 4. An algorithm of the AR model for imputation

Since both general and rare patterns are of interest in the process of imputation, it is reasonable to choose a sufficiently small value of support. It has been found experimentally that $s_{min}=0.01$ is optimal for finding rare patterns. The minimum confidence is defined on the basis of a relative frequency of the most common value (MCV) in the attribute. This is because replacement of a missing value with the most common one is a priori more effective than the use of a rule with confidence less than the relative frequency of the MCV. So the value of c_{min} is defined for every dataset individually. We do not apply restrictions concerning maximum support and confidence since both nontrivial and obvious patterns are suitable for imputation.

The following three models are based on machine learning algorithms – a random forest, a support vector machine, and a neural network – and, therefore, they need preliminary setting of training and test sets. To describe data, we introduce the following notations. Let $D=(D^1, D^2, ..., \text{and } D^j)$ be an incomplete data matrix, and let $j=(\overline{1,k})$ be a range of attributes that may contain missing values. To predict the missing values by using one of the above-mentioned algorithms for each attribute D^j containing missing values, we can divide D into four parts:

– observed values of the attribute D^j, denoted as y^j_{obs};
– missing values of the attribute D^j, denoted as y^j_{mis};
– values of the other attributes, which correspond to the observed values of the attribute D^j, denoted as x^j_{obs};
– values of the other attributes, which correspond to the missing values of the attribute D^j, denoted as x^j_{mis}.

For imputation of missing values in each attribute j, we train one of the algorithms on the complete parts of the dataset x^j_{obs} and y^j_{obs}, where y^j_{obs} represents class labels. To impute the missing values y^j_{mis} of an attribute, we use a training model for prediction of class labels based on x^j_{mis}.

2. A model on the basis of a support vector machine (SVM). The main parameter determining its work is the kernel function *k_funct*. In order to take into account probable non-linear relationships between the attributes, we will use the radial basis function (RBF) as a kernel function that has the least number of settings affecting the algorithm usefulness in comparison with other kernels and requires fewer mathematical calculations. Since the support vector machine works only with numerical data, we use non-numeric attributes only as class labels, excluding them from the training and the test sets x^j_{obs} and y^j_{obs}. The algorithm of the SVM model looks as follows (Fig. 5).

Algorithm 3. Imputation of missing values on the basis of a support vector machine:
Input: $k \times n$ of the matrix D and the kernel function ***k_funct***
1: Clustering of D;
2: $s \leftarrow$ the number of clusters;
3: ***for q*** *in* ***s do***:
4: Partition of D_q into $[y_{obs}^j]_q$, $[y_{mis}^j]_q$, $[x_{obs}^j]_q$, and $[x_{mis}^j]_q$;
5: Normalization of $[y_{obs}^j]_q$, $[y_{mis}^j]_q$, $[x_{obs}^j]_q$, and $[x_{mis}^j]_q$;
6: $k_1 \leftarrow$ the number of attributes with omissions;
7: ***for j*** *in* k_1 ***do***:
8: Construction of optimal hyperplanes separating class labels $[y_{obs}^j]_q$ with the use of the kernel function ***k_funct***;
9: Maximizing the margin $\|w\|$ between the hyperplanes;
10: Prediction of $[y_{mis}^j]_q$ based on $[x_{mis}^j]_q$;
11: $D_{imp}^j \leftarrow$ imputation of missing values for the attribute ***j***;
12: ***end for***
13: ***end for***
14: ***return*** matrix with imputed values D^{imp}.

Fig. 5. An algorithm of the SVM model for imputation

3. A model on the basis of a random forest (RF). The main parameters determining its work are the number of trees in the forest c and the number of features to consider when splitting a node p. The algorithm of the RF model looks as follows (Fig. 6).

Algorithm 4. Imputation of missing values on the basis of a random forest:
Input: $k \times n$ of the matrix D, the number of trees in the forest c, and the number of features to consider when splitting the node p.
1: Clustering of D;
2: $s \leftarrow$ the number of clusters;
3: ***for q*** *in* ***s do***:
4: Partition of D_q into $[y_{obs}^j]_q$, $[y_{mis}^j]_q$, $[x_{obs}^j]_q$, and $[x_{mis}^j]_q$;
5: $k_1 \leftarrow$ the number of attributes with omissions;
6: ***for j*** *in* k_1 ***do***:
7: Building of a random forest with specified parameters $\{h([x_{obs}^j]_q, \Theta_k=p), k=1, ...\}_{i=1}^{N=c}$ and training on $[y_{obs}^j]_q \sim [x_{obs}^j]_q$;
8: Prediction of $[y_{mis}^j]_q$ based on $[x_{mis}^j]_q$;
9: $D_{imp}^j \leftarrow$ imputation of missing values for the attribute ***j***;
10: ***end for***
11: ***end for***
12: ***return*** matrix with imputed values D^{imp}.

Fig. 6. An algorithm of the RF model for imputation

4. A model on the basis of an artificial neural network. Studies show that neural networks are successfully used for imputation of missing data. A multilayer perceptron (MLP) with one or two hidden layers and a sigmoid

function as an activation function is one of the most widespread basic architectures for the process of imputation. The number of neurons in the input layer is equal to the number of input attributes, and the number of neurons in the output layer corresponds to the number of unique values of the attribute that is being classified. The training process is carried out with the observed data in a batch mode with the back-propagation learning algorithm. The main parameters determining its work are the activation function *act_funct*, the number of hidden layers h, the number of neurons N, and the number of iterations (epochs) e. The algorithm of the MLP model looks as follows (Fig. 7).

Algorithm 5. Imputation of missing values on the basis of a neural network:
Input: $k \times n$ of the matrix D, the activation function ***act_funct***, the number of hidden layers h, the number of neurons N, and the number of iterations e.
1: Clustering of D;
2: $s \leftarrow$ the number of clusters;
3: ***for*** q *in* s ***do***:
4: Partition of D_q into $[y_{obs}^j]_q$, $[y_{mis}^j]_q$, $[x_{obs}^j]_q$, and $[x_{mis}^j]_q$;
5: Normalization of $[y_{obs}^j]_q$, $[y_{mis}^j]_q$, $[x_{obs}^j]_q$, and $[x_{mis}^j]_q$;
6: $k_1 \leftarrow$ the number of attributes with omissions;
7: ***for*** j *in* k_1 ***do***:
8: Building of a multilayer perceptron ***MLP***(act_funct, h, N) with the activation function ***act_funct***, hidden layers h, and neurons N;
9: Training on $[y_{obs}^j]_q$~$[x_{obs}^j]_q$ over e iterations;
10: Prediction of $[y_{mis}^j]_q$ based on $[x_{mis}^j]_q$;
11: $D_{imp}^j \leftarrow$ imputation of missing values for the attribute j;
12: ***end for***
13: ***end for***
14: ***return*** matrix with imputed values D^{imp}.

Fig. 7. An algorithm of the MLP model for imputation

5. A model on the basis of an EM algorithm (EM). To describe the algorithm work, let us introduce the following notation. Suppose we have a statistical model involving a vector of observed data X_{obs}, a vector of missing values X_{mis}, and a vector of unknown parameters θ, then there is the log likelihood function $L(\theta; X_{obs}, X_{mis})=p(X_{obs}, X_{mis}|\theta)$. The maximum likelihood estimate of the unknown parameters is determined by the marginal likelihood of the observed data $L(\theta; X_{obs})=p(X_{obs}|\theta)$. The EM algorithm tries to find the maximum likelihood estimate of the marginal likelihood by iteratively applying the following two steps: expectation (an E-step) and maximization (an M-step). In the E-step, the expected value of the log likelihood function $Q(\theta|\theta^{(t)})$ is calculated under the current estimate of the parameters $\theta^{(t)}$. In the M-step $\theta^{(t+1)}$, $Q(\theta^{(t+1)}; \theta^{(t)}) \geq \geq Q(\theta; \theta^{(t)})$. The E-step and the M-step are repeated alternately until the difference $L(\theta^{(t+1)})-L(\theta^{(t)})$ is less than *thr*, where *thr* is a convergence limit [23]. However, the EM algorithm works only with numerical data, so it can not handle missing values in nominal attributes. The main parameters that determine its work are the maximum number of iterations for an expectation maximization i and the convergence limit *thr*. The algorithm of the EM model looks as follows (Fig. 8).

The algorithms of models for imputation (Fig. 4–8) use the proposed approach on the basis of the extended matrix of attributes when forming a training dataset. The problem of a large number of unique values in the attributes is solved by the step of preliminary clustering before the models' training. In order to assess the performance of the proposed models for imputation, we study their work on the same dataset with different numbers of simulated missing values.

Algorithm 6. Imputation of missing values on the basis of the EM algorithm:
Input: $k \times n$ of the matrix D, the maximum number of iterations i, and the convergence limit ***thr***.
1: Clustering of D;
2: $s \leftarrow$ the number of clusters;
3: ***for*** q *in* s ***do***:
4: $k_1 \leftarrow$ the number of attributes with omissions;
5: ***for*** j *in* k_1 ***do***:
6: ***num_it*** $\leftarrow 0$;
7: ***while num_it*** $\leq i$ ***do:***
8: Calculation of the expected value of the log-likelihood $Q(\theta|\theta^{(t)})$ under the current estimate of the parameters $\theta^{(t)}$;
9: Computation of the $\theta^{(t+1)}$;
10: ***if*** $L(\theta^{(t+1)}) - L(\theta^{(t)}) \leq thr$ ***then break***;
11: ***else num_it*** ++;
12: $D_{imp}^j \leftarrow$ imputation of missing values for the attribute j;
13: ***end for***
14: ***end for***
15: ***return*** matrix with imputed values D^{imp}.

Fig. 8. An algorithm of the EM model for imputation

5. A study of the imputation models' performance and a design of the imputation models' ensembles

5. 1. A study of missing data estimation errors

To test the performance of the proposed models, we use the above-mentioned real datasets from social network accounts. To get an impartial assessment, we form the complete model dataset M′ by deletion of missing values. Furthermore, we randomly generate in M′ different numbers of simulated omissions (1, 2, 5, 10, 20, 30, 50, and 70 %) and apply the suggested models to the obtained datasets. Thus, we can impartially assess the performance of the proposed models since the real values that are deleted synthetically are known. We offer the following approaches to estimate the errors of the imputation. The first one is applicable both to nominal and numerical variables. The second one is based on the root mean squared error and is suitable only for numerical variables.

Let I_{inp} be a vector of imputed values of an attribute K, with I_{real} being a vector of real data and n_I being the number of values in these vectors. Let us designate the frequency of incorrectly imputed values as n_{I_false} and the frequency of correctly imputed values as n_{I_true}, so that $n_{I\_false}+n_{I\_true}=n_I$. To assess the imputation performance while allowing for a deviation of imputed values from the real ones in some range Δr, we can define the frequency of the correctly imputed values as follows:

$$n_{I\_true} = \sum_{i=1}^{n_I} I_{imp} \in \left[I_{real_i} - \Delta r; I_{real_i} + \Delta r \right]. \quad (3)$$

The imputation error based on the frequency of the correctly imputed values is determined in the following way:

$$\delta = \frac{n_{I\_true}}{n_I} \cdot 100\,\%. \quad (4)$$

Assessment of imputation by using the root mean squared error is widely made in various studies [14, 24]. It is applicable only to numerical indicators and has a large number of modifications; however, the most popular variant is based on the normalized root mean squared error (NRMSE) [25], which is defined as follows:

$$NRMSE = \sqrt{\frac{mean\left(\left(I_{imp} - I_{real}\right)^2\right)}{var\left(I_{real}\right)}}, \quad (5)$$

where I_{imp} is a vector of the imputed values, I_{real} is a vector of the real data, mean is a sample mean of the squared difference between the values of the real and imputed vectors, var is the variance of the real values. Values of NRMSE close to 0 indicate good performance of the imputation and closeness between the imputed and real vectors, and conversely, the closer the values of the NRMSE are to 1, the lower the performance is.

Using the error estimates described above, let us assess the performance of the proposed models of imputation. We also compare them with one of the most popular and widely used methods, called the most common value (MCV), which are usually integrated into statistical software and platforms.

As we can see from Fig. 1, the following four attributes contain missing values: V1 (age), V2 (marital status), V3 (year of graduation from university), and V4 (year of leaving school); three of them are numerical (V1, V3, and V4), but one of them is nominal (V2).

Fig. 9 shows the imputation results of the nominal attribute V2 based on δ. Proceeding from the nominal nature of V2, the EM model of imputation is not applied, and the deviation of the imputed values from the real ones is not allowed; therefore, Δr=0.

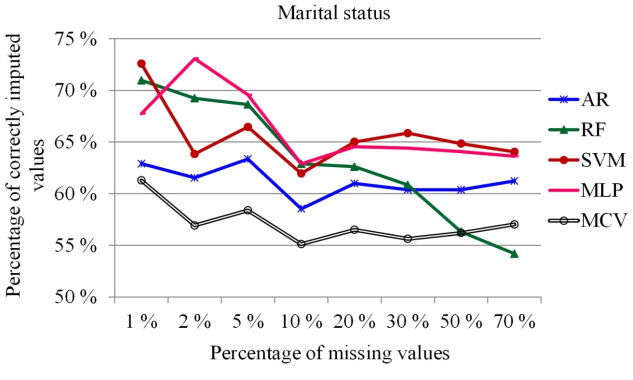


Fig. 9. Performance of the proposed models based on the δ error when imputing the attribute V2: AR is a model on the basis of association rules, RF is a model on the basis of a random forest, SVM is a model on the basis of a support vector machine, MLP is a model on the basis of a neural network, and MCV is a method of imputation of the "Most Common Value"

Thus, we can see that the MCV performs worse than the other models. In general, the MLP allows obtaining slightly better results in comparison with the AR, the RF and the SVM because it performs stably well for the whole percentage of incomplete data. The RF also performs well when the rate of missing values is low, but it worsens with their growing number. The SVM is slightly inferior to the RF and the MLP, given a small percentage of omissions, but as this percentage increases, the SVM performs on par with the MLP.

Let us consider the performance of the models in case of imputation of numeric data while using the attribute V1 as an example. Fig. 10 shows the performance of the five proposed models and the MCV method.

We can see from the values of the δ error of a numerical attribute that the AR and the RF perform better than the other models both within Δr=0 (Fig. 10, *a*) and Δr=1 (Fig. 10, *b*). However, when the rate of missing values is more than 30 %, the EM outperforms the AR and the RF by about 10 to 15 %. At the same time, the MCV is substantially inferior to the other models, which indicates its inapplicability for imputation of data from social networks' accounts. Let us compare the performance of the proposed models by using the normalized root mean squared error NRMSE (Fig. 11).

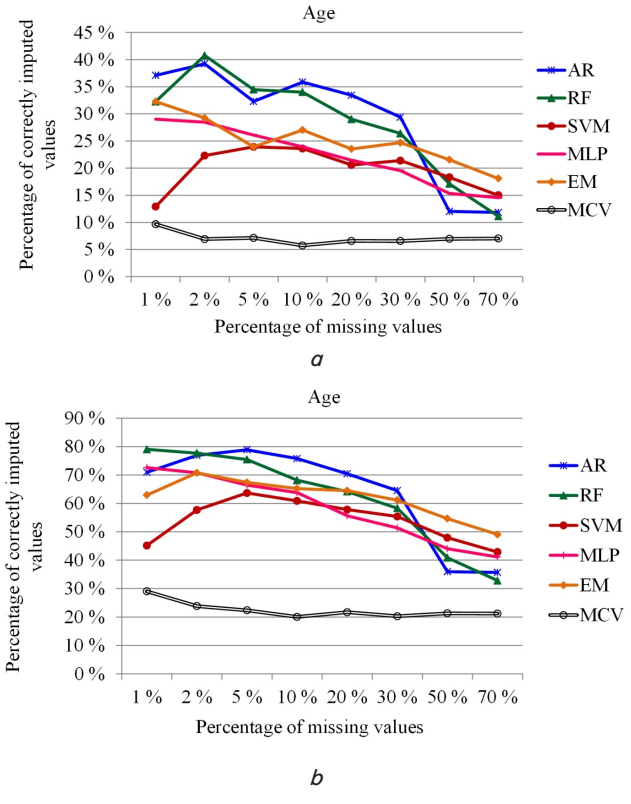


Fig. 10. Performance of the proposed models based on the δ error when imputing the attribute V1: AR is a model on the basis of association rules, RF is a model on the basis of a random forest, SVM is a model on the basis of a support vector machine, MLP is a model on the basis of a neural network, EM is a model on the basis of the EM algorithm, MCV is a method of imputation of the "Most Common Value": *a* is within a range of Δr=0; *b* is within a range of Δr=1

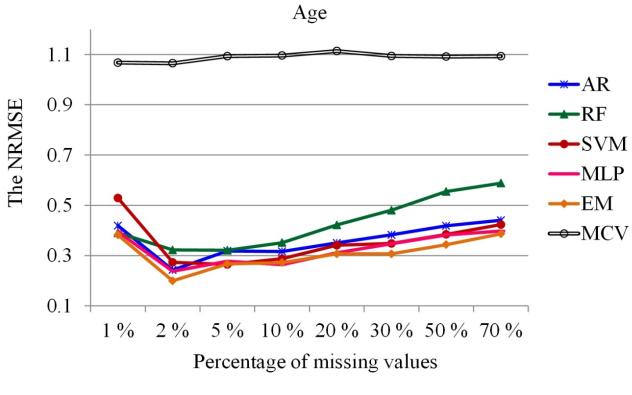


Fig. 11. The NRMSE when imputing the attribute V1

As we can see from Fig. 11, the NRMSE of the MCV also indicates its inapplicability to this type of data. The EM, the MLP and the SVM perform better than the other models, showing the lowest values of the NRMSE and a

gradual increase in the error with the growing number of missing values. The AR has slightly worse values of the NRMSE than the EM, the MLP and the SVM (by 0.04-0.08), which do not tend to increase substantially with an increase in the percentage of omissions. The RF performs the worst in respect to the other models since it has the largest values of the NRMSE and a substantial increase in the error with an increase in the rate of missing values.

5. 2. A design of ensembles of the imputation models

To improve the quality and robustness of imputation, we develop ensembles of the imputation models for numeric and nominal data on the basis of the models discussed above. Based on the heterogeneity of the imputation models' errors when imputing the categorical attribute, we take into account the performance of each model. To do it, we assign some weights to these models' outputs according to the expert ranking based on their δ errors (Fig. 9) as follows: the MLP – 0.29, the SVM – 0.26, the RF – 0.24, and the AR – 0.21. We do not consider the MCV method since it has the poorest performance. The structure of the proposed ensemble of the models ECAT (an ensemble for categorical attributes) is shown in Fig. 12.

Based on the performance analysis of the imputation models on numeric data, we cannot select any model that outperforms the others in terms of δ and NRMSE errors. The AR and the RF have the highest percentages of correctly imputed values of δ and are inferior to the SVM, the MLP and the EM, with higher values of the NRMSE. And conversely, the SVM, the MLP and the EM produce smaller values of the NRMSE but have much lower levels of correctly imputed values in terms of the δ error. Therefore, it was decided to construct the ensemble ENUM (an ensemble for numeric attributes), using all the five models (Fig. 13).

Since none of them can be marked out as a better one than others, the unweighed voting is used when combining the results. We implement the following two approaches to combining the ensembles' outputs: outputs averaging (ENUM_A) and unweighed majority voting (ENUM_V).

6. Discussion of the performance of the proposed ensembles

To study the performance and robustness of the ECAT ensemble, we compare it with the single models. The performance of the proposed ensemble is shown in Fig. 14.

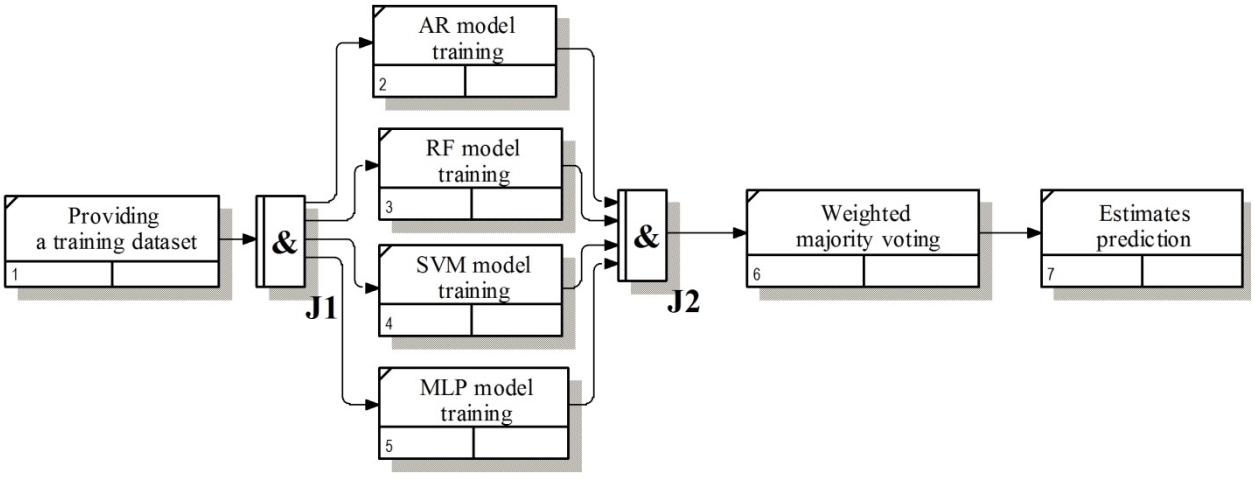


Fig. 12. The structure of the ECAT ensemble for imputation of nominal data

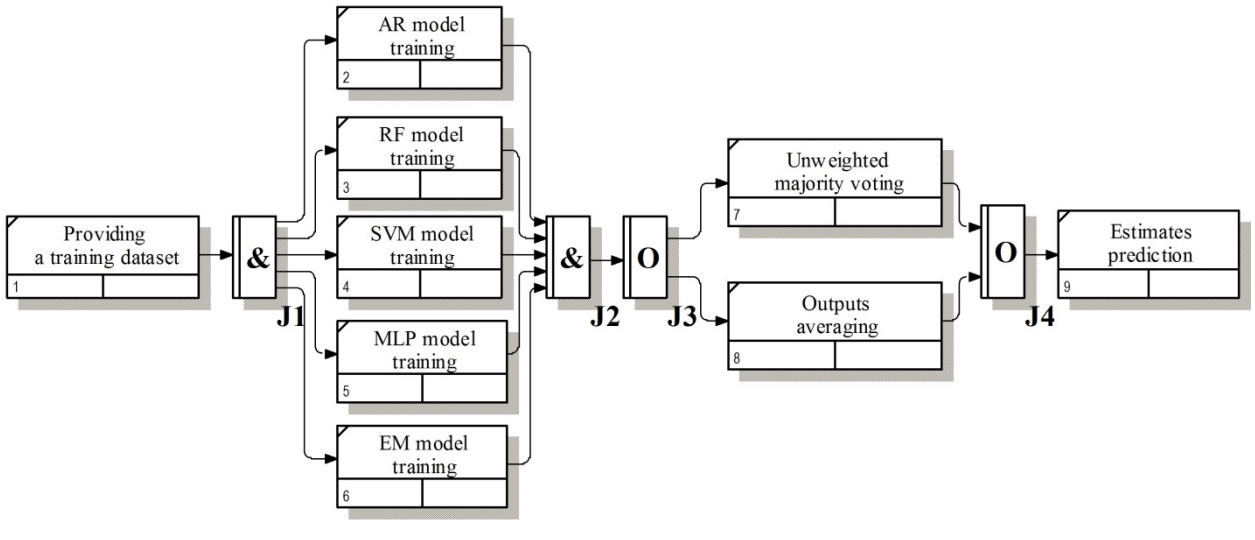


Fig. 13. The structure of the ENUM_A and ENUM_V ensembles for imputation of nominal data

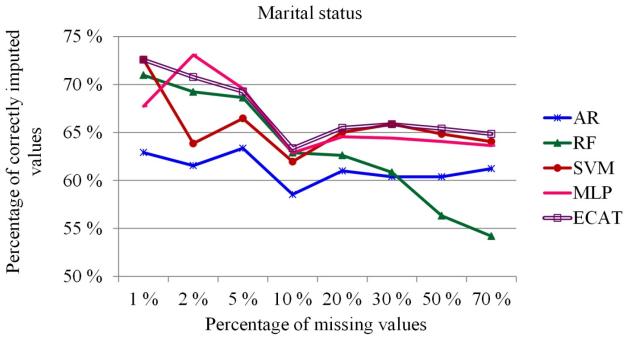
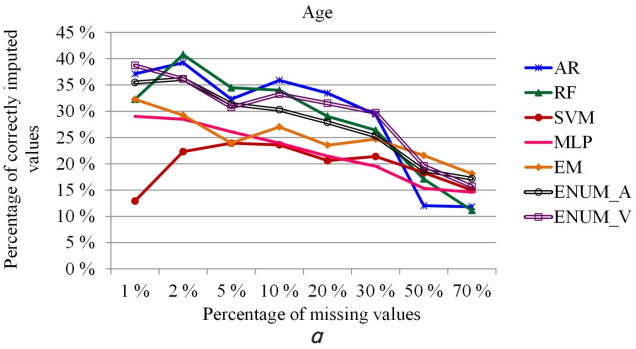


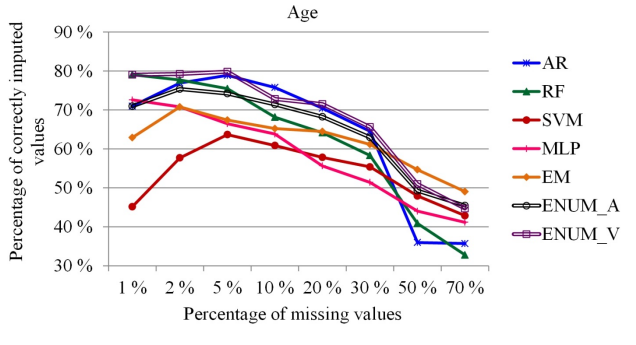
Fig. 14. Performance of the ECAT ensemble compared to the single models based on the δ error when imputing the attribute V2

As we can see from Fig. 14, the ECAT helps increase the robustness of imputation in comparison with the single models. The ECAT has steadily high percentage of the correctly imputed values for all rates of omissions, and in general it outperforms the AR, the RF, the SVM, and the MLP.

Fig. 15 shows the performance of the constructed ensemble ENUM with different combinations of outputs of the ENUM_A and the ENUM_V based on the imputation error δ compared to the single models. As we can see from the obtained values of δ, the ENUM_A and the ENUM_V are on par with the AR and the RF and outperform the other single models in the presence of up to 30 % of missing values. When the rate of omissions is more or equal to 30 %, the ENUM_A and the ENUM_V even tend to outperform the AR and the RF.



a



b

Fig. 15. Performance of the single models and the ensembles ENUM_A and ENUM_V based on the δ error when imputing the attribute V1: *a* is within a range of Δr=0; *b* is within a range of Δr=1

Thus, the proposed ensembles of the models make it possible to obtain more robust results of imputation compared to the single models. Comparing the ENUM_A and the ENUM_V, it can be seen that their performances are practically equal. The ENUM_V slightly outperforms the ENUM_A both at Δr=0 and within the allowable deviation of Δr=1.

Let us compare the performance of the proposed ensembles in terms of the normalized root mean squared error NRMSE (Fig. 16).

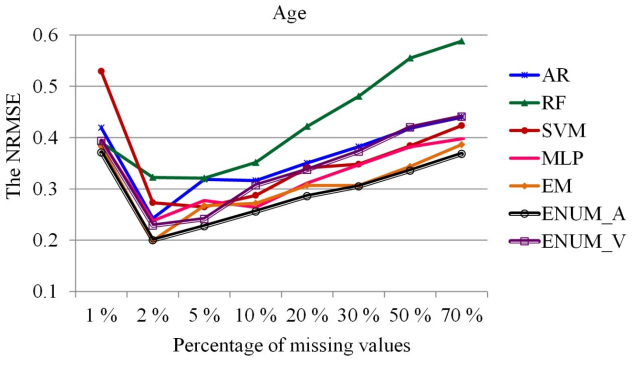


Fig. 16. The NRMSE of single models and the ensembles ENUM_A and ENUM_V when imputing the attribute V1

As we can clearly see from Fig. 16, the ENUM_A produces the smallest values of the NRMSE at any rate of omissions and thus outperforms the other single models. The ENUM_V has slightly more values of the error that are on par with the SVM, the MLP, and the EM. At the same time, the ENUM_V significantly outperforms the AR in the presence of a low rate of missing values and the RF for any percentage of omissions.

Thus, the performance evaluation of the proposed ensembles demonstrates that they are able to impute correctly numeric data on par with the best single models AR and RF in terms of the imputation error δ while producing the NRMSE that is less or equal to this error in the best single models SVM, MLP, and EM. Therefore, the ensembles of the models allow obtaining such results that are practically "the golden mean" in terms of both errors, since they have satisfactory values of both δ and NRMSE errors.

Introduction of preliminary clustering has been found conducive to overcoming the problem of a large number of unique values in attributes when designing models for imputation of missing data from social networks. Creating of a training set on the basis of the extended matrix of attributes allows adding a range of informative complete attributes that correlate with incomplete data of the incomplete matrix and thus making initial data more informative in further designing of improved methods for imputation. The use of the proposed ensembles of the models makes it possible to obtain more robust results in comparison with single models in terms of δ and NRMSE errors by means of the diversity of the base models.

However, we should note that parametric identification of weighting coefficients of models' outputs is not automatic; it requires involvement of an analyst when combining outputs of an ensemble with weighed voting. Furthermore, the performance of the proposed models is not considered in terms of computational complexity of the used algorithms

since the main purpose of this study is to maximize the correctness of imputation of the δ error and minimize the root mean squared error NRMSE.

The designed models and ensembles of imputation help improve the quality of initial data from social networks when pre-processing and preparing data for further analysis by analysts of social networks. The suggested models of imputation can be used by analysts and programmers when developing, organizing and supporting social network services for users in order to increase effectiveness and stability of their activities through the imputation of lost or missing data.

Previously published studies of authors in the area of improving the quality of data from social networks are further developed in this work. We have complemented the models on the basis of association rules and a random forest by a range of models on the basis of a support vector machine, a neural network, and an EM algorithm. We have improved the process of creating a training dataset for the models through forming an extended matrix of attributes and designing ensembles of the models. Further development of the results of this work can be carried out in the following directions: analysis of the proposed models in terms of computational complexity of the used algorithms and development of a method for automatic parametric identification of weighting coefficients of the models' outputs for an ensemble with weighed voting, for example, on the basis of training errors of the models. Moreover, an important prospective issue is to develop an information technology for imputation of missing data that is able to automate the process of data pre-processing for further analysis.

7. Conclusions

1. Based on the structure of real data from social network accounts, the study has shown that they can be divided into two types: those that potentially contain missing values and correspond to a matrix of incomplete data X_1 and those that always have complete data, which correspond to an enrichment matrix X_2. The research of the structure and strength of correlations between the attributes of X_1 and X_2 confirms a statistically significant relationship. The existence of the correlation justifies the use of the extended matrix of attributes $D=X_1+X_2$ and thus makes initial data more informative in further designing of improved methods for imputation.

2. Taking into consideration the complex nature of social network data, we have proposed imputation models using the following algorithms: association rules, a random forest, a support vector machine, a multilayer perceptron, and an EM algorithm. In view of the revealed correlation, the problem of a large number of unique values in attributes can be overcome by an approach on the basis of preliminary clustering being introduced into the models of imputation. We have shown that this approach makes it possible to reduce the number of unique values in potentially incomplete attributes inside the obtained clusters by 40–65 %. To assess the performance of the models, the following two errors were used: a relative error δ and a normalized root mean squared error NRMSE. We have compared the performance of the proposed models with the popular method of imputation MCV being integrated into statistical packages and demonstrated its inapplicability of maintaining this type of data from social network accounts.

3. To improve the quality and robustness of imputation, we have designed ensembles of the models for imputation of nominal (ECAT) and numeric attributes (ENUM_A and ENUM_V) on the basis of the suggested models. Combinations of the base-model outputs are based on the two imputation errors – δ and NRMSE. We have shown that the designed ensembles help increase efficiency and robustness of imputation, namely they improve the correctness of imputation by up to 9 % in comparison with the average results of single models and decrease the NRMSE error on the average by 0.1.

References

1. Slabchenko, O. O. Uluchsheniie kachestva iskhodnykh dannykh v zadachakh modelirovaniia internet-soobshchestv na osnove kompleksnogo primeneniia modelei segmentatsii, imputatsii i obogashcheniya dannykh [Text] / O. O. Slabchenko, V. N. Sydorenko // Visnik KrNU. – 2013. – Issue 6. – P. 50–58.
2. Kossinets, G. Effects of missing data in social networks [Text] / G. Kossinets // Social networks. – 2006. – Vol. 28, Issue 3. – P. 247–268. doi: 10.1016/j.socnet.2005.07.002
3. Nakagawa, S. Missing inaction: the dangers of ignoring missing data [Text] / S. Nakagawa, R. P. Freckleton // Trends in Ecology and Evolution. – 2008. – Vol. 23, Issue 11. – P. 592–596. doi: 10.1016/j.tree.2008.06.01
4. Graham, J. W. Missing Data Analysis: Making It Work in the Real World [Text] / J. W. Graham // Annual Review of Psychology. – 2009. – Vol. 60, Issue 1. – P. 549–576. doi: 10.1146/annurev.psych.58.110405.085530
5. Rhoads, C. H. Problems with Tests of the Missingness Mechanism in Quantitative Policy Studies [Text] / C. H. Rhoads // Statistics, Politics, and Policy. – 2012. – Vol. 3, Issue 1. doi: 10.1515/2151-7509.1012
6. Schafer, J. L. Missing data: our view of the state of the art [Text] / J. L. Schafer, J. W. Graham // Psychological Methods. – 2002. – Vol. 7, Issue 2. – P. 147–177. doi: 10.1037/1082-989x.7.2.147
7. Schlomer, G. L. Best Practices for Missing Data Management in Counseling Psychology [Text] / G. L. Schlomer, S. Bauman, N. A. Card // Journal of Counseling Psychology. – 2010. – Vol. 57, Issue 1. – P. 1–10. doi: 10.1037/a0018082
8. Chang, C. Weighting Adjustments in Survey Sampling [Text] / C. Chang, F. B. Butar // European International Journal of Science and Technology. – 2013. – Vol. 2, Issue 9. – P. 214–236.
9. Huisman, M. Imputation of missing network data: Some simple procedures [Text] / M. Huisman // Journal of Social Structure. – 2009. – Vol. 10, Issue 1. – P. 1–10.

10. Baraldi, A. N. An introduction to modern missing data analyses [Text] / A. N. Baraldi, C. K. Enders // Journal of School Psychology. – 2010. – Vol. 48, Issue 1. – P. 5–37. doi: 10.1016/j.jsp.2009.10.001
11. Andridge, R. R. A Review of Hot Deck Imputation for Survey Non-response [Text] / R. R. Andridge, R. J. A. Little // International Statistical Review. – 2010. – Vol. 78, Issue 1. – P. 40–64. doi: 10.1111/j.1751-5823.2010.00103.x
12. Schmitt, P. A Comparison of Six Methods for Missing Data Imputation [Text] / P. Schmitt, J. Mandel, M. Guedj // Journal of Biometrics & Biostatistics. – 2015. – Vol. 06, Issue 01. doi: 10.4172/2155-6180.1000224
13. Silva-Ramirez, E.-L. Missing value imputation on missing completely at random data using multilayer perceptrons [Text] / E.-L. Silva-Ramírez, R. Pino-Mejías, M. López-Coello, M.-D. Cubiles-de-la-Vega // Neural Networks. – 2011. – Vol. 24, Issue 1. – P. 121–129. doi: 10.1016/j.neunet.2010.09.008
14. Aydilek, I. B. A hybrid method for imputation of missing values using optimized fuzzy c-means with support vector regression and a genetic algorithm [Text] / I. B. Aydilek, A. Arslan // Information Sciences. – 2013. – Vol. 233. – P. 25–35. doi: 10.1016/j.ins.2013.01.021
15. Jerez, J. M. Missing data imputation using statistical and machine learning methods in a real breast cancer problem [Text] / J. M. Jerez, I. Molina, P. J. García-Laencina, E. Alba, N. Ribelles, M. Martín, L. Franco // Artificial Intelligence in Medicine. – 2010. – Vol. 50, Issue 2. – P. 105–115. doi: 10.1016/j.artmed.2010.05.002
16. Nanni, L. A classifier ensemble approach for the missing feature problem [Text] / L. Nanni, A. Lumini, S. Brahnam // Artificial Intelligence in Medicine. – 2012. – Vol. 55, Issue 1. – P. 37–50. doi: 10.1016/j.artmed.2011.11.006
17. Gebregziabher, M. Latent class based multiple imputation approach for missing categorical data [Text] / M. Gebregziabher, S. M. DeSantis // Journal of Statistical Planning and Inference. – 2010. – Vol. 140, Issue 11. – P. 3252–3262. doi: 10.1016/j.jspi.2010.04.020
18. Senapti, R. A novel approach for missing value imputation and classification of microarray dataset [Text] / R. Senapti, K. Shaw, S. Mishra, D. Mishra // Procedia Engineering. – 2012. – Vol. 38. – P. 1067–1071. doi: 10.1016/j.proeng.2012.06.134
19. Stekhoven, D. J. MissForest – non-parametric missing value imputation for mixed-type data [Text] / D. J. Stekhoven, P. Buehlmann // Bioinformatics. – 2011. – Vol. 28, Issue 1. – P. 112–118. doi: 10.1093/bioinformatics/btr597
20. Gheyas, I. A. A neural network-based framework for the reconstruction of incomplete data sets [Text] / I. A. Gheyas, L. S. Smith // Neurocomputing. – 2010. – Vol. 73, Issue 16-18. – P. 3039–3065. doi: 10.1016/j.neucom.2010.06.021
21. Slabchenko, O. Analysis and synthesis of models on basis of machine learning for missing values imputation from social networks' personal accounts [Text] / O. Slabchenko, V. Sydorenko // Visnik KrNU. – 2014. – Issue 5. – P. 105–111.
22. Chekmyshev, O. A. Izvlecheniie i ispolzovaniie dannykh iz elektronnykh sotsialnykh setei [Text] / O. A. Chekmyshev, A. D. Yashunskiy. – Moscow: IPM im. M. V. Keldysha, 2014. – 16 p.
23. Little, J. A. Statistical Analysis with Missing Data [Text] / J. A. Little, D. B. Rubin. – 2-nd ed. – New Jersey: John Wiley & Sons, 2002. – 408 p. doi: 10.1002/9781119013563
24. Ferrari, P. A. An imputation method for categorical variables with application to nonlinear principal component analysis [Text] / P. A. Ferrari, P. Annoni, A. Barbiero, G. Manzi // Computational Statistics & Data Analysis. – 2011. – Vol. 55, Issue 7. – P. 2410–2420. doi: 10.1016/j.csda.2011.02.007
25. Oba, S. A Bayesian missing value estimation method for gene expression profile data [Text] / S. Oba, M.-a. Sato, I. Takemasa, M. Monden, K.-I. Matsubara, S. Ishii // Bioinformatics. – 2003. – Vol. 19, Issue 16. – P. 2088–2096. doi: 10.1093/bioinformatics/btg287