UDC 001.891.57, 004.8+303.72

ANALYSIS AND SYNTHESIS OF MODELS ON BASIS OF MACHINE LEARNING FOR MISSING VALUES IMPUTATION FROM SOCIAL NETWORKS' PERSONAL ACCOUNTS

O. Slabchenko, V. Sydorenko

Kremenchuk Mykhailo Ostrohradskiy National University

vul. Pershotravneva, 20, Kremenchuk, 39600, Ukraine. E-mail: vnsidorenko@gmail.com

Problem statement of missing values imputation from the personal accounts of social network users was performed. A structure flow chart of data imputation model was offered, which includes the following machine learning algorithms: association rules, decision trees and random forests. Missing values imputation experiment with use of the offered models was carried out on the full dataset and in a particular cluster. An interval estimate of imputation accuracy within the range from ± 0 (point estimate) to ± 2 was performed. It was ascertained that the preliminary clustering improves the performance quality of the imputation model based on association rules, while the models based on decision trees and random forests are practically insensitive to it. It was shown that the best results of missing values imputation procedure show the models based on decision trees and random forests.

Key words: social network analysis, data imputation, machine learning, association rules, decision trees, random forests.

АНАЛІЗ І СИНТЕЗ МОДЕЛЕЙ НА ОСНОВІ МАШИННОГО НАВЧАННЯ ДЛЯ ІМПУТАЦІЇ ДАНИХ ІЗ ПЕРСОНАЛЬНИХ АКАУНТІВ СОЦІАЛЬНИХ МЕРЕЖ

О. О. Слабченко, В. М. Сидоренко

Кременчуцький національний університет імені Михайла Остроградського

вул. Першотравнева, 20, м. Кременчук, 39600, Україна. E-mail: slabchenko.olesia@gmail.com

Виконано постановку задачі імпутації даних з пропущеними значеннями з персональних акаунтів користувачів соціальних мереж. Запропоновано базову структурну схему моделі імпутації некомплектних даних, яка включає в себе наступні алгоритми машинного навчання: асоціативні правила, дерева та ліси рішень. Проведено експеримент з відновлення пропущених значень з використанням запропонованих моделей на повній множині даних і в окремому сегменті. Виконано інтервальну оцінку правильності відновлення пропусків у межах від ±0 (точкова оцінка) до ±2. Встановлено, що попередня кластеризація підвищує якість роботи моделі імпутації на основі асоціативних правил, тоді як моделі на основі дерев і лісів рішень практично нечутливі до неї. Показано, що найкращі результати відновлення пропущених значень показують моделі на основі дерев і лісів рішень.

Ключові слова: соціально-мережевий аналіз, імпутація даних, машинне навчання, асоціативні правила, дерева рішень, ліси рішень.

INTRODUCTION. Missing data are inevitable problem when real data processing [1]. They are extremely undesirable when applying data mining algorithms [2], because they result the reduction of efficiency. Furthermore, most data mining algorithms cannot work directly with incomplete data. Traditional and most popular methods of missing data processing are: ignoring, deleting, substitution [3–6]. However, they cause the loss of efficiency and getting biased estimates due to deletion of incomplete cases [7]. Moreover, this problem worsens when processing high dimensional data. So to get simulation results more qualitative and reliable the improvement of initial data quality is necessary. It is possible with use of more complex and modern model-based methods of missing data imputation [1, 7, 8].

Recent trends in the progress of methods for missing values imputation show that is extremely difficult to develop a universal model that could show good results in various subject areas. Therefore, in many papers models are mainly offered for specific domains: psychology, medicine [9], biology [10], genetics [11], social science [12], software cost estimation [13], databases [14].

Our research is aimed at improving the quality of data stored in social networks through analysis and imputation of missing values that may be contained in the personal user accounts in the context of improving the results of models application in social network analysis. However, it is very difficult to process such kind of data because they have both numerical and non-numerical nature, and contain a large number of contradictions.

Another feature of the data stored in social networks is the existence of connections between them. Investigation of the structure of these ties provides additional data describing users and their activity on the network [15]. In addition, they can store the information about latent factors which cause the formation of connections between accounts that are similar to each other in some way, and combining them into a group. Thus, the data from social networks contain not only the missing values, but also the information for their imputation or enrichment represented in implicit or explicit form.

In view of the fact that a large number of methods were developed to solve the problem of incomplete data, the analysis and synthesis of optimal missing values imputation models with use of known algorithms is urgent prob-

lem for this domain.

The goal of this work is the analysis and synthesis of data imputation models as applied to the personal user accounts of social networks, taking into account the mixed nature of their attributes on basis of machine learning models.

MATERIAL AND RESEARCH RESULTS. Let there be some initial matrix Z that has the dimensionality $k \times n$, elements z_{ij} ($i = \overline{1,n}$, $j = \overline{1,k}$) of which describe n users in space of k quantitative and qualitative indicators.

Since the objects in a social network are connected, then the ties between them can be represented as a simple undirected graph $G = (V, E)$ where V is a nonempty set of vertices, E is a set of vertices' pairs which are called edges. Connection between the vertices of the graph can be defined by the adjacency matrix A of the dimension $n \times n$, where $a_{ij} = 1$, if there is a connection between the vertices of the graph G with numbers i and j, and $a_{ij} = 0$ otherwise.

On the other hand, the specific of data (attributes) from the social networks accounts is that the qualitative-quantitative indicators describing a user can be divided into 2 groups: those which may contain missing values, and attributes which are always complete (for example, time of the last browsing a page or privacy settings). Let denote the matrix of incomplete data as X_1 which has the dimensionality $n \times k_1$, where k_1 is the number of indicators that may contain missing values. X_1 has the following form: $X_1 = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1k_1} \\ x_{21} & x_{22} & \dots & x_{2k_1} \\ \dots & \dots & \dots & \dots \\ x_{n1} & x_{n2} & \dots & x_{nk_1} \end{bmatrix}$. The matrix of always complete data is denoted as X_2 and has the dimensionality $n \times k_2$, where k_2 is the number of complete indicators. Matrix X_2 will be called the enrichment matrix. It has the following form: $X_2 = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1k_2} \\ x_{21} & x_{22} & \dots & x_{2k_2} \\ \dots & \dots & \dots & \dots \\ x_{n1} & x_{n2} & \dots & x_{nk_2} \end{bmatrix}$. Then the matrix D can be represented as the junction of X_1 and X_2:

$$D = \left[\begin{array}{cccc|cccc} x_{11} & x_{12} & \dots & x_{1k_1} & x_{1k_1+1} & x_{1k_1+2} & \dots & x_{1k} \\ x_{21} & x_{22} & \dots & x_{2k_1} & x_{2k_1+1} & x_{2k_1+2} & \dots & x_{2k} \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ x_{n1} & x_{n2} & \dots & x_{nk_1} & x_{nk_1+1} & x_{nk_1+2} & \dots & x_{nk} \end{array}\right],$$

where $k = k_1 + k_2$.

Let introduce the concept of the complete data matrix X and the binary matrix M of the dimension $k \times n$, wherein $m_{ij} = 1$ if x_{ij} is not a missing and $m_{ij} = "NA"$ if the corresponding element x_{ij} is missing. Then the matrix with missing values X_1 can be represented as $X_1 = X \circ\circ M$, where the operator "$\circ\circ$" is defined by analogy with the Hadamard product "$\circ$", which means the element wise matrix-matrix multiplication of X and M.

That is: $\begin{cases} x_{ij} \circ\circ m_{ij} = x_{ij} \cdot m_{ij} = x_{ij} & if \quad m_{ij} = 1 \\ x_{ij} \circ\circ m_{ij} = "NA" & if \quad m_{ij} = "NA" \end{cases}$.

For instance:

$$X_1 = X \circ\circ M = \begin{pmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \\ x_{31} & x_{32} \end{pmatrix} \circ\circ \begin{pmatrix} 1 & "NA" \\ "NA" & 1 \\ 1 & "NA" \end{pmatrix} =$$

$$= \begin{pmatrix} x_{11} & "NA" \\ "NA" & x_{21} \\ x_{31} & "NA" \end{pmatrix}.$$

Using the introduced terminology, the problem statement of imputation can be formulated as follows: the incomplete matrix X_1 is specified; it required to restore the original matrix X, on the basis of the data that are contained in the incomplete matrix X_1 and the enrichment matrix X_2. Given that completing matrix X with absolutely accurate values is extremely difficult, it is possible to find a reasonable estimate $\hat{X} \approx X$, using the share of incorrectly imputed values as a performance criterion and allowing deviations of values derived from initial data in a specified range Δr.

However, taking into account that the user profiles, whose indicators require imputation, have real connections that are defined by the adjacency matrix A, the imputation problem can be reformulated as: given a triple (X_1, X_2, A); it is required to restore the original matrix X or find its reasonable estimate $\hat{X} \approx X$, on the basis of not only the information contained in the incomplete matrix X_1, but also the additional matrix X_2 and the adjacency matrix A.

In this paper, missing values imputation experiments were performed on the data collected from personal user accounts from the city of Kremenchug in the social network "Vkontakte". Totally the information on about 200,000 accounts containing missing values was obtained.

Attributes that are available for analysis, the following: age, relationship status, years of school and university graduation, the number of subscribers, subscriptions and friends, sex, privacy settings, contacts outside of the social network, the number of downloaded media files, time and date of the last browsing the page. Let define the set P which has the cardinality of k_1 variables that may contain missing values and therefore are target for imputation: P_{k_1} = {age; relationship

status; year of a university graduation; year of a school graduation}. Since the matrix X_2 is involved in the subsequent calculations, and provides a source of additional information when restoring the values of X_1, apart from the $k_1 = 4$ target for the imputation $k_2 = 8$ extra variables also take part in the calculations.

A brief outline of the data preparation to carry out the comparative modelling looks as following:

1. Collecting data about users from the city of Kremenchug and forming a model matrix Z that has the dimensionality $k \times n$ by deleting all missing data, where $n = 2774$ is the number of records and k is the number of indicators.

2. Generating in Z 1, 2, 5, 10, 20, 30, 50 and 70 % of missing values by turns that have uniform discrete distribution, and specifying them as "missing", for every attribute. Getting eight incomplete matrices Z_i', $i = \overline{1,8}$.

After the study of these data it was ascertained that the main problem, which complicates the analysis, is their mixed nature. Therefore, the imputation of such data requires the algorithms allowing simultaneous processing of numerical and categorical indicators. On the assumption of this criterion, an attempt to use the models on basis of association rules, decision trees and random forests was undertaken for missing values imputation in data. Fig. 1 represents the offered missing data imputation model. The block "imputation model" corresponds to one of three models: association rules, decision tree or random forest, which are analysed by turns.

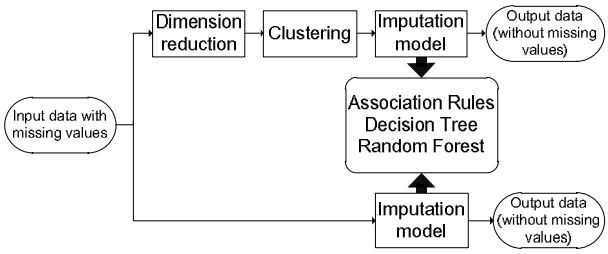

Рисунок 1 – Flowchart of the offered model for imputing missing data

In case of association rules' use two types of models were built: the first with the algorithm of building the rules for all data [16] from Z_i', the other one with use of improved algorithm, which was offered by authors in [17], and includes the stage of preliminary clustering. The basis of the preliminary clustering is a considerable number of unique values, which is typical for every attribute and makes it difficult or impossible to find rare but significant rules.

The models of random forests and decision trees were built with use of CART algorithm, and had the inputs with identical parameters. These models are limited by depth in order to prevent overfitting. The random forests consist of $n\_estimators = 100$ trees. The models on basis of decision trees and random forests also were built for the two cases: of all data and of certain clusters.

Obviously, it is difficult to analyse the 5 input informative indicators, so there is the problem of reducing their dimensions. Its solution is possible by using of the methods based on search and detection of correlations between variables and further identifying of latent factors, which describe and consolidate similar variables, such as factor analysis.

The goal of any factor analysis method is to represent the matrix element of input data Z as a linear combination of several factors. Then, the matrix element z_{ij} can be expressed as a linear combination of r factors [18]:

$$z_{ij} = y_{i1}p_{1j} + y_{i2}p_{2j} + \ldots + y_{ir}p_{rj} \qquad (1)$$

Equation (1) expresses the basic model of factor analysis and can be written in matrix form:

$$Z = YP, \qquad (2)$$

where Z is the matrix of the input data. Y is the matrix, which needs to be determined; Y is called a factor pattern, and its coefficients are called factor loadings. P is a matrix of values of all factors for all individuals. At that, Y shows the relationship between the variables and factors, P describes certain individuals. For the correlation matrix R of the input data Z he following relation takes place:

$$\frac{1}{n-1}ZZ' = R, \qquad (3)$$

where $\frac{1}{n-1}$ is a scalar. When substituting (3) in (2) it is obtained $R = \frac{1}{n-1}ZZ' = A\frac{1}{n-1}PP'A'$. Expression $\frac{1}{n-1}PP' = \tilde{N}$ is the correlation matrix, which reflects the relationships between factors:

$$R = ACA' \qquad (4)$$

When imposing the condition of uncorrelated factors $(C = 1)$ in this equation it is obtained:

$$R = AA' \qquad (5)$$

Equations (4) and (5) are called the fundamental theorem of factor analysis. C – is the matrix of correlation coefficients between the factors. In the case where the orthogonal factors are postulated, C becomes an identity matrix and it is omitted when multiplying [18].

In our case the orthogonality of desired factors is assumed. The factor analysis was performed by applying the principal component method [19] and the further procedure of Varimax rotation. This procedure maximizes the dispersion of loadings' squares for the each factor, which leads to increase of large and decrease of low values of the factor loadings. The number

of factors is determined based on the criterion of Cattell [20]. Table 1 shows the percentage of the variables variance explained by the obtained factors.

Table 1 – The percentage of variables' variance explained by factors

Factor Number	Eigenvalue	Percent of Variance	Cumulative Percentage
1	2,5958	54,583	54,583
2	1,23913	26,056	80,639
3	0,550029	11,566	92,205
4	0,170861	3,593	95,797
5	0,127169	2,674	98,471
6	0,0691975	1,455	99,927
7	0,00349454	0,073	100,000

As can be seen from Table 1 the obtained orthogonal factors explain 92,2 % of the variables' variance. Table 2 shows the factor loadings matrix after the rotation of axes.

Table 2 – Factor loading matrix after Varimax rotation

Variables	Factor 1	Factor 2	Factor 3
age	–0,888021	–0,0277293	0,0347097
relation code	0,0418745	0,057447	–0,20032
UG	0,803159	0,0001681	0,0017988
SG	0,956649	0,0461541	–0,0897558
followers	0,131469	0,32324	0,32443
subscriptions	0,0455038	0,643792	–0,108763
friends	0,0709865	0,57298	0,199264
sex	0,0934906	0,0284965	0,0225422
openness	–0,143864	–0,0785016	–0,0639811
activity	0,224398	0,275299	0,622552
communicativeness	0,24438	0,203178	0,0634779
last seen	0,0268603	0,0958093	0,529805

Based on Table 2, the obtained factors were personified as follows: age characteristic (1st factor, 54,583 % of the variance), communicativeness (2nd factor, 26,056 % of the variance) and online activity (the third factor, 11,566 % of the variance). Thus, the results of the factor analysis have shown that for the description of social network users it is possible to pass from space of $k = 12$ correlated indicators to three-dimensional space of derived factors.

Taking into account the efficiency of the computational procedure for large sample sizes and the numerical nature of the analysed variables, clustering is performed with use of the k – means method with the Euclidean metric and calculation of the squared Euclidean distance as a disparity measure. K – means method belongs to the class of algorithms which require a prior setting of the number of desired clusters.

One of the methods allowing to find the optimal number of clusters is the Elbow criterion. It looks at the percentage of variance explained as a function of the number of clusters and chooses the number so that adding a new cluster does not improve the results of simulation [21]. Percentage of variance explained is the ratio of the sample variance within the cluster to the value of the total sample variance.

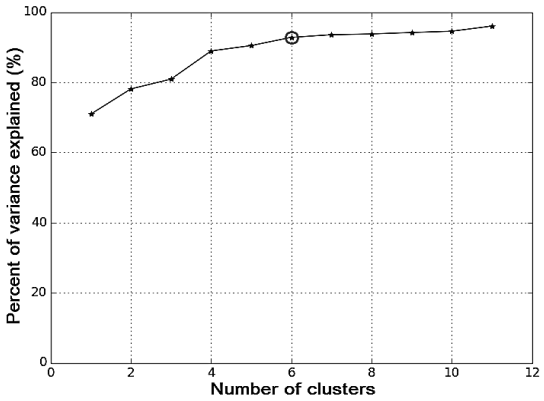

Figure 2 – Percentage of variance explained by the clusters on the number of clusters

As can be seen from Fig. 2, the first clusters explain a significant amount of the variance, but at some point this increase stops. The number of clusters is chosen at this point is equal to 6 and satisfies the elbow criterion. The results of the clustering procedure are shown in Fig. 3–4. The resulting clusters have the following structure: 1st contains 402 records, 2nd – 85, 3rd –777, 4th – 608, 5th – 229, and 6th – 673.

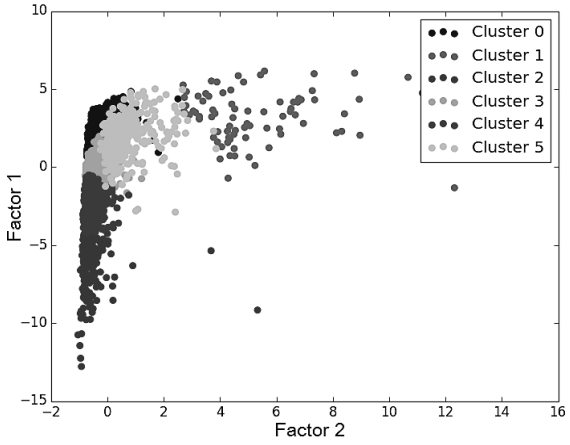

Figure 3 – The results of clustering in the space of 1 and 2 factors

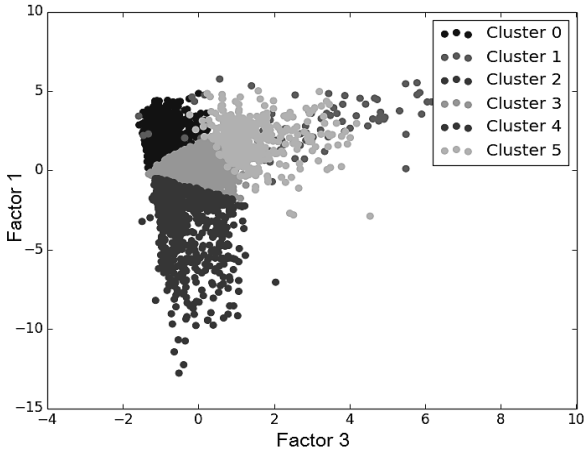

Figure 4 – The results of clustering in the space of 1 and 3 factors

Further in accordance to the offered flow chart, 2 types of imputation model were built: for the all dataset and within clusters. Furthermore, for getting more stable results $j = 5$ sets of incomplete data were generated for each of $i = \overline{1,8}$ percentage of missing values. Thus, j experiments were performed for each of the models, and their results were averaged.

As an example, the comparative results of the imputation of variables "age" and "year of a school graduation" are shown in Fig. 5–7 for all data and within the largest third cluster containing 777 records. An interval estimate of missing values imputation accuracy was performed for numerical variables (age, year of a university graduation etc.) within ± 0 (exact accuracy, curves without any markers), also for $\Delta r = \pm 1$ (curves with diamond-shaped markers) and $\Delta r = \pm 2$ (curves with triangular markers), since users are usually classified within a certain group.

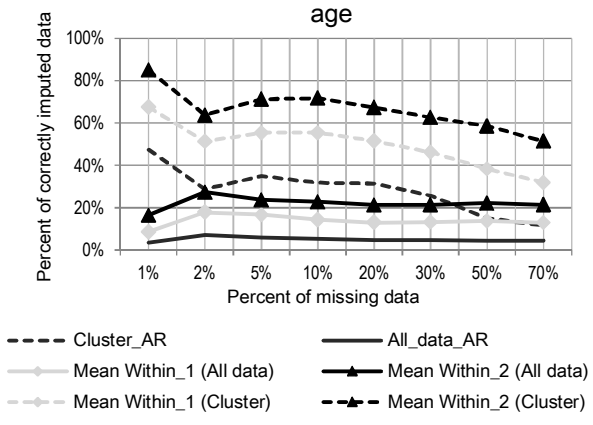

Figure 5 – The results of the model's work on basis of association rules

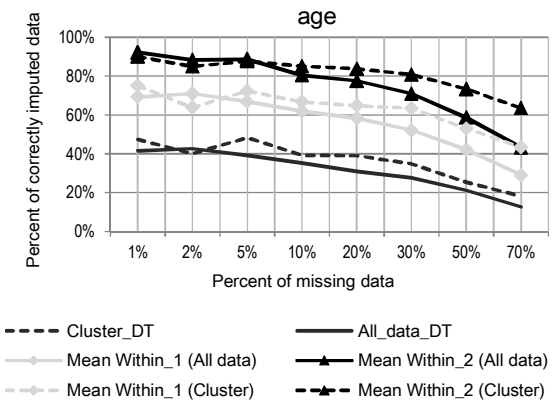

Figure 6 – The results of the model's work on basis of decision trees

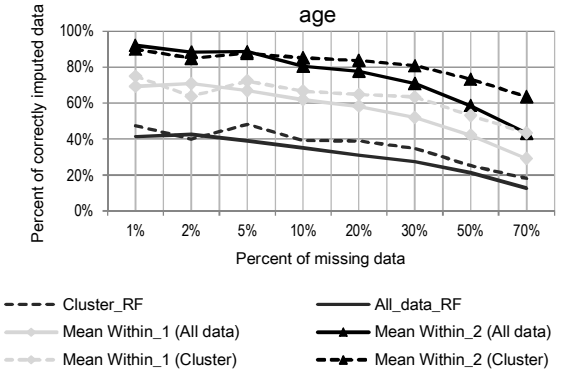

Figure 7 – The results of the model's work on basis of random forest

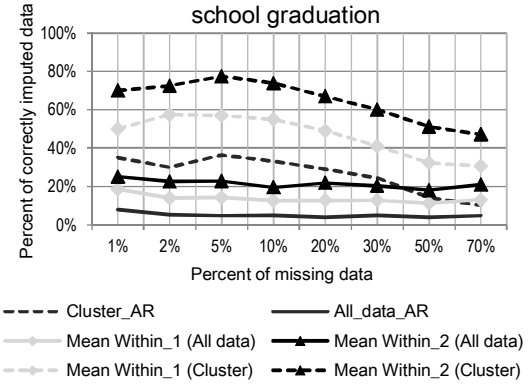

Figure 8 – The results of the model's work on basis of association rules

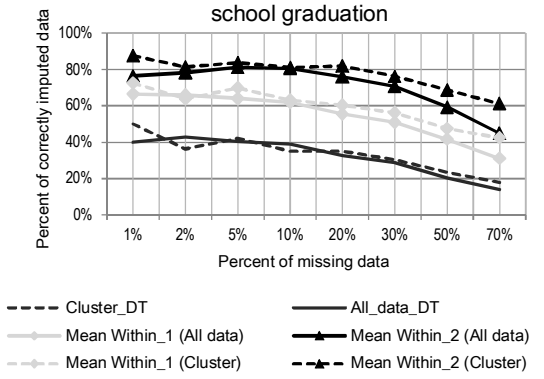

Figure 9 – The results of the model's work on basis of decision trees

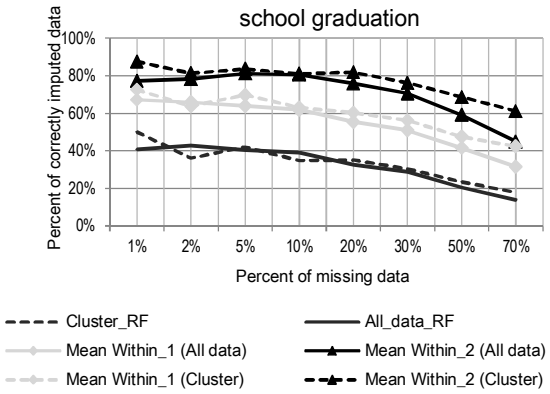

Figure 11 – The results of the model's work on basis of random forest

As it can be seen from the figures, the model on basis of association rules imputes missing values with less accuracy than the models on basis of decision trees and random forests, and shows sensitivity to the procedure of pre-clustering. The models based on decision trees and random forests, show nearly the same results of the imputation, what can be caused by specifying identical parameters when their training. Therefore, Fig. 12, 13 present the comparative results of the models on basis of association rules and decision trees with the exact accuracy and within the intervals $\Delta r = 1$ and $\Delta r = 2$ in the cluster.

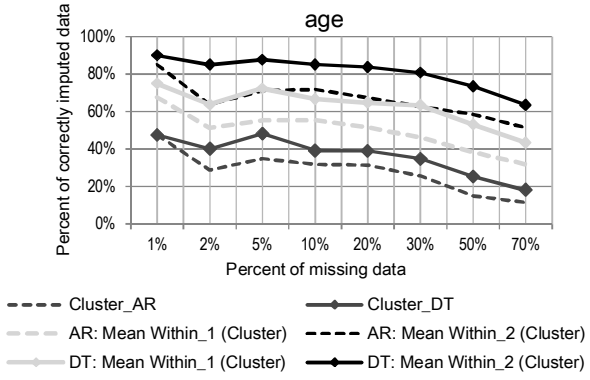

Figure 12 – Comparative results of the models' work on basis of association rules and decision trees (variable "Age")

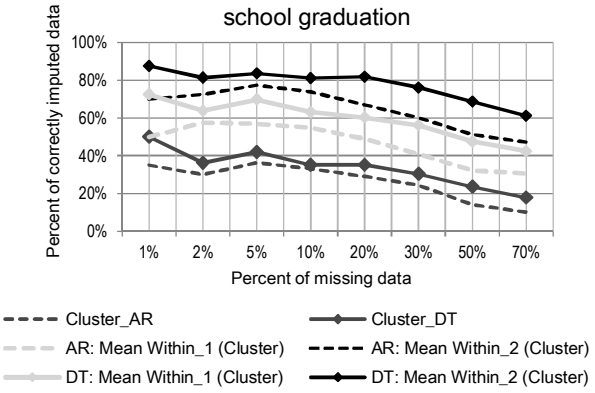

Figure 13 – Comparative results of the models' work on basis of association rules and decision trees (variable "Year of a school graduation")

The dotted curves represent the results of the model based on association rules, solid curves represent the results of the model based on decision trees. The lower group of curves shows the results of missing values imputation with exact accuracy, the light-coloured curves in the middle of the plot represent imputation results within the range ± 1, and the dark curves on the top of the plot represent imputation results within the range ± 2.

CONCLUSION. As experiment results show, models on basis of association rules impute missing values with less accuracy in the range from ± 0 to ± 2, than models on basis of trees and random forests, and show sensitivity to the procedure of preliminary clustering: accuracy of data imputation is higher in the cluster from 20 to 50 % than on the whole data set. Models on basis of trees and random forests show almost identical results of imputation, which may be caused by specifying the same parameters while training. They are practically insensitive to the pre-clustering, stably demonstrate high accuracy of imputation within ± 1 (80–50 %) and ± 2 (90–60 %) with up to 50 % of missing values. From the comparative results of the models on basis of association rules and decision trees it can be seen that the last one shows better results (on the average the accuracy of imputation is higher from 5 to 15 %).

Improvement of the results of a model based on association rules is possible by implementation of the rules' weights that can be computed on basis of explicit and implicit connections analysis between users in a social network.

Since the training quality of the models on basis of decision trees and random forests depends on the input parameters, the accuracy of missing values imputation by these models can be improved by optimal parameters selection when training, for example, with use of genetic algorithms.

REFERENCES

1. Little, R.J. and Rubin, D.B. (1990), *Statistich-eskiy analiz dannykh s propuskami* [Statistical analysis with missing data], Finansy i statistika, Moscow, Russia.

2. Aydilek, I.B., and Arslan, A. (2013), "A hybrid method for imputation of missing values using optimized fuzzy c-means with support vector regression and a genetic algorithm", *Information Sciences*, vol. 233, pp. 25–35.

3. Baraldi, A.N, and Enders, C.K. (2009), "An introduction to modern missing data analyses", *Journal of School Psychology*, vol. 48, no. 1, pp. 5–37.

4. Paul, A.D. (2001), *Missing Data (Quantitative Applications in the Social Sciences)*, SAGE Publications, USA.

5. Quinten, A., and Raaijmakers, Q.A.W. (1999), Effectiveness of different missing data treatments in surveys with Likert-type data: Introducing the relative

mean substitution approach, *Educational and Psychological Measurement*, vol. 59, pp. 725–748.

6. Graham, J. W. (2012), *Missing Data: Analysis and Design*, Springer New York Heidelberg Dordrecht London.

7. Little, R.J.A., and Rubin, D.B. (2002), *Statistical analysis with missing data,* second ed., Wiley, Canada.

8. Gheyas, I. A., and Smith, L. S. (2010), "A neural network-based framework for the reconstruction of incomplete data sets", *Neurocomputing*, vol. 73, no. 16–18, pp. 3039–3065.

9. Nanni, L., Lumini, A., and Brahnam, S. (2012), "A classifier ensemble approach for the missing feature problem", *Artificial Intelligence in Medicine*, vol. 55, no. 1, pp. 37–50.

10. Gebregziabher, M., and DeSantis, S.M. (2010), "Latent class based multiple imputation approach for missing categorical data", *Journal of Statistical Planning and Inference*, vol. 140, no. 11, pp. 3252–3262.

11. Senapti, R., Shaw, K., Mishra, S., and Mishra, D. (2012), "A novel approach for missing value imputation and classification of microarray dataset", *Procedia Engineering*, vol. 38, pp. 1067–1071.

12. Silva-Ramirez, E.L., Pino-Mejias, R., Lopez-Coello, M., and Cubiles-de-la-Vega, M.A. (2011), Missing value imputation on missing completely at random data using multilayer perceptrons, *Neural Networks*, vol. 24, no. 1, pp. 121–129.

13. Sentas, P., and Angelis, L. (2006), Categorical missing data imputation for software cost estimation by multinomial logistic regression, *Journal of Systems and Software*, vol. 79, no. 3, pp. 404–414.

14. Abdella, M., and Marwala, T. The use of genetic algorithms and neural networks to approximate missing data in database // *Proceedings of the 3rd International Conference on Computational Cybernetics*, April 13–16, Mauritius. – USA, 2005. – PP. 207–212.

15. Dorogovtsev, S. and Mendes, J. (2002), "Evolution of networks", *Advances in Physics*, vol. 51, no. 4, pp. 1079–1187.

16. Kaiser, J. (2011), "Algorithm for Missing Values Imputation in Categorical Data with Use of Association Rules", *ACEEE International Journal on Recent Trends in Engineering & Technology*, vol. 6, no. 1, pp. 111–114.

17. Slabchenko, O.O., and Sydorenko, V.N. (2014) "An improved algorithm for imputation data from social network accounts with use of association rules" // *Materialy XXI Mezhdunarodnoy nauchno-prakticheskoy konferentsii studentov, aspirantov i molodykh uchenykh KrNU imeni Mikhaila Ostrogradskogo "Aktual'nyye problemy zhiznedeyatel'nosti obshchestva"* [Proceedings of the 21 International scientific and practical conference of students, PhD students and young scientists of Kremenchuk Mykhailo Ostrohradskiy National University on Actual problems of society activity], Kremenchug, April 24–25, 2014.

18. Iberla, K. (1980), *Faktornyy analiz* [Factor analysis], M: Statistica.

19. Ayvazyan, S.A., Bukhshtaber, V.M., Yenyukov, I.S., and Meshalkin, L.D. (1989), *Prikladnaya statistika: Klassifikatsiya i snizheniye razmernosti* [Applied statistics: Classification and dimensionality reduction], M: Finansy i statistica.

20. Kim, D.O., and M'yuller, C.U. (1989), Faktornyy analiz: statisticheskiye metody i prakticheskiye voprosy [Factor analysis: statistical methods and practical issues], M: Finansy i statistica.

21. See, E.G., Ketchen, D.J., Shook, J.L., and Shook, C.L. (1996), "The application of cluster analysis in Strategic Management Research: An analysis and critique", *Strategic Management Journal*, vol. 17, no. 6, pp. 441–458.

АНАЛИЗ И СИНТЕЗ МОДЕЛЕЙ НА ОСНОВЕ МАШИННОГО ОБУЧЕНИЯ ДЛЯ ИМПУТАЦИИ ДАННЫХ ИЗ ПЕРСОНАЛЬНЫХ АККАУНТОВ СОЦИАЛЬНЫХ СЕТЕЙ

О. О. Слабченко, В. Н. Сидоренко

Кременчугский национальный университет имени Михаила Остроградского

ул. Первомайская, 20, г. Кременчуг, 39600, Украина. E-mail: slabchenko.olesia@gmail.com

Выполнена постановка задачи импутации данных с пропущенными значениями из персональных аккаунтов пользователей социальных сетей. Предложена базовая структурная схема модели импутации некомплектных данных, включающая следующие алгоритмы машинного обучения: ассоциативные правила, деревья и леса решений. Проведен эксперимент по восстановлению пропущенных значений с использованием предложенных моделей на полном множестве данных и в отдельном сегменте. Выполнена интервальная оценка правильности восстановления пропусков в пределах от ±0 (точечная оценка) до ±2. Установлено, что предварительная сегментация повышает качество работы модели импутации на основе ассоциативных правил, в то время как модели на основе деревьев и лесов решений практически нечувствительны к ней. Показано, что наилучшие результаты восстановления пропущенных значений показывают модели на основе деревьев и лесов решений.

Ключевые слова: социально-сетевой анализ, импутация данных, машинное обучение, ассоциативные правила, деревья решений, леса решений.

Стаття надійшла 14.09.2014.