Mathematical Problems of Computer Science 40, 55--67, 2013.

Hierarchical Cluster Analysis for Partially Synthetic Data Generation

Levon H. Aslanyan and Vardan H. Topchyan

Institute for Informatics and Automation Problems of NAS RA
e-mail: lasl@sci.am, vardan.topchyan@gmail.com

Abstract

Limiting the risk of information disclosure is now common for statistical agencies. One of the widespread approaches is to release the synthetic, public use of microdata sets. To put it another way, thanks to the multiple imputations the sensitive variables of original data are replaced by new/synthetic values. This paper introduces the method for partially synthetic data generation based on hierarchical cluster analysis.

Keywords: Confidentiality, Multiple imputation, Synthetic data, Hierarchical clustering.

1. Introduction

Submission of sociological and/or economic data to the public structures is an integral part of procedures in statistical organizations. However, this task should not assume the risk of disclosure of sensitive or personal information. Analysis of the published data in this area [1] indicates the presence of diverse approaches/methods for solving such problems, including variable recoding, swapping data, and adding noise values. Although these methods replace the original data, protecting the information in this way may lead to distortion of relationships between the different segments of the data set, which in turn can lead to erroneous conclusions/inferences on the stage of data analysis, such as methods of standard statistical processing.

An alternative approach to solve this problem, which also tries to maintain functional relationships between the segments of the data set simultaneously, is the approach of fully synthetic data generation [2]. In this case, the statistical organization should: (i) randomly and independently record the general format and content of critical information units, as well as integrate them into the corresponding set of expected synthetic data; (ii) establish new/synthetic values in the information units by the selected strategy; (iii) provide a number of generated synthetic data sets to the public. There are various methods [3]-[5] for generating fully synthetic data providing the receipt of meaningful results using standard statistical methods.

In spite of advantages of the fully synthetic data, the process of generating these data is quite time-consuming. In this regard, statistical organizations often use partially synthetic data

which is a mix of original and synthesized data [6]. For example, the statistical agency may seek to protect the confidentiality of certain records, or may not allow the identification of several records. In connection with that, the synthetic values generated only for certain variables and the values of the other variables remain changed.

As in the case of fully synthetic data, partially synthetic data are also providing the restriction of the disclosure risk, allowing to obtain meaningful results by using standard statistical analysis. Note that, due to its nature, the use of partially synthetic data provide more accurate statistical calculations. For the same reason, the risk of disclosure is higher than in case of the fully synthetic data. However, there are several algorithms [7]-[9] for generating partially synthetic data which are used by many statistical agencies (US Federal Reserve Board, US Bureau of the Census, etc.) that indicate the perspectives of this method.

The above mentioned situation was considered as a basis for the analysis of non-parametric methods for generating partially synthetic data sets used for calculation of simple estimands (average, standard deviation, etc.) and for construction of data driven linear regression models [3]. Published literature [10] shows that one of the most known non-parametric approaches is the method of sequential imputations of variables [11]. [11] uses CART (*Classification and Regression Trees*) model [12] for this reason. In this case, in non-parametric method [11] the hierarchical clustering model can be assumed to substitute CART as an alternative approach.

The paper is organized as follows. Section 2 reviews the description of partially synthetic data, the principle of their generation and analysis. Section 3 illustrates how the hierarchical clustering model can be used in a similar to CART way, for generating partially synthetic data. Section 4 concludes the discussion about using hierarchical clustering in partially synthetic data generation.

2. Partially Synthetic Data

2.1 Creation of Partially Synthetic Data

For partially synthetic data definition, we use notation [13]. The process of generating partially synthetic data consists of two parts: (i) pretreatment/preprocessing of data; (ii) replacement of the corresponding/tagged values with the synthetic one. Formally, this process can be described as follows.

Let U be the set of records/information units, $U = \{U_1, U_2, ..., U_N\}$, where each information unit $U_i (1 \le i \le N)$ is characterized by the p attributes/ variables, $Y = \{Y_1, Y_2, ..., Y_p\}$.

$$U = \begin{array}{c|cccc} & Y_1 & Y_2 & \Lambda & Y_p \\ U_1 & y_{11} & y_{12} & \Lambda & y_{1p} \\ U_2 & y_{21} & y_{22} & \Lambda & y_{2p} \\ & & & & \\ \text{M} & & & & \\ & & & & \\ U_N & y_{N1} & y_{N2} & \Lambda & y_{Np} \end{array}$$

During preprocessing information units and confidential variables (rows and columns of matrix U) are selected, and the threshold conditions for these variables are set.

Let $n(n \le N)$ be the number of randomly selected information units that will be considered in the current observation, denote those units by $\{U_{i_1}, U_{i_2}, ..., U_{i_n}\}$.

Similarly, $d(d \le p)$ is the number of confidential variables, $\{Y_{j_1}, Y_{j_2}, ..., Y_{j_d}\}$. In addition, N-block $I = (I_1, I_2, ..., I_N)$ and p-block $J = (J_1, J_2, ..., J_p)$ of numbers are defined as follows:

$$I_r = \begin{cases} 1, U_r \in \{U_{i_1}, U_{i_2}, ..., U_{i_n}\} \\ 0, U_r \notin \{U_{i_1}, U_{i_2}, ..., U_{i_n}\} \end{cases}, \qquad 1 \le r \le N,$$

$$J_k = \begin{cases} 1, Y_k \in \{Y_{j_1}, Y_{j_2}, ..., Y_{j_d}\} \\ 0, Y_k \notin \{Y_{j_1}, Y_{j_2}, ..., Y_{j_d}\} \end{cases}, \qquad 1 \le k \le p.$$

The process of determining the information units and confidential variables is presented in the following scheme.

$$U = \begin{array}{c|c|cccccc} & & J_1 & \Lambda & \Lambda & \Lambda & J_p \\ & & Y_1 & \Lambda & \Lambda & \Lambda & Y_p \\ \hline I_1 & U_1 & y_{11} & & & & y_{1p} \\ \mathbb{N} & \mathbb{N} & & & & & \\ \mathbb{N} & \mathbb{N} & & & & & \\ \mathbb{N} & \mathbb{N} & & & & & \\ \mathbb{N} & \mathbb{N} & & & & & \\ I_N & U_N & y_{N1} & & & & y_{Np} \end{array}$$

As a result, $U_{obs} = (U_{rep}, U_{nrep})$ matrix is defined. Here U_{rep} is a $[n \times d]$ matrix with the values of confidential variables $\{Y_{j_1}, Y_{j_2}, ..., Y_{j_d}\}$ and U_{nrep} $[n \times (p-d)]$ is a matrix of other variables values (replaced vs. not replaced).

$$U_{obs} = \begin{array}{c|ccccc} & \multicolumn{3}{c}{U_{rep}} & & U_{nrep} \\ & Y_{j_1} & \Lambda & Y_{j_d} & \Lambda & Y_{j_p} \\ \hline U_{i_1} & y_{i_1 j_1} & \Lambda & y_{i_1 j_d} & \Lambda & y_{i_1 j_p} \\ U_{i_2} & y_{i_2 j_1} & \Lambda & y_{i_2 j_d} & \Lambda & y_{i_2 j_p} \\ \mathbb{M} & & & & & \\ U_{i_r} & y_{i_r j_1} & \Lambda & y_{i_r j_d} & \Lambda & y_{i_r j_p} \\ \mathbb{M} & & & & & \\ U_{i_n} & y_{i_n j_1} & \Lambda & y_{i_n j_d} & \Lambda & y_{i_n j_p} \end{array}$$

Next, the $c_{j_1}, c_{j_2}, ..., c_{j_d}$ threshold conditions are set for the $Y_{j_1}, Y_{j_2}, ..., Y_{j_d}$ variables.

$$U_{rep} = \begin{array}{c|ccc} & C_{j_1} & \Lambda & C_{j_d} \\ & Y_{j_1} & \Lambda & Y_{j_d} \\ \hline U_{i_1} & y_{i_1 j_1} & \Lambda & y_{i_1 j_d} \\ U_{i_2} & y_{i_2 j_1} & \Lambda & y_{i_2 j_d} \\ \mathbb{M} & & & \\ U_{i_r} & y_{i_r j_1} & \Lambda & y_{i_r j_d} \\ \mathbb{M} & & & \\ U_{i_n} & y_{i_n j_1} & \Lambda & y_{i_n j_d} \end{array}$$

Based on these conditions, the indicator matrix $Z[n \times d]$ is defined.

$$Z = \begin{array}{c|ccc} & z_{1 j_1} & \Lambda & z_{1 j_d} \\ & z_{2 j_1} & \Lambda & z_{2 j_d} \\ \mathbb{M} & & & \\ & z_{r j_1} & \Lambda & z_{r j_d} \\ \mathbb{M} & & & \\ & z_{n j_1} & \Lambda & z_{n j_d} \end{array}$$

This matrix describes the need to replace the corresponding values of the confidential variables.

Thus, during preprocessing $U_{obs} = (U_{rep}, U_{nrep})$ and Z matrices are defined. The resulting, observed data set are denoted as $D = (U_{rep}, U_{nrep}, Z)$.

The second part of the partially synthetic data generation is the process of replacement. Namely, based on $D = (U_{rep}, U_{nrep}, Z)$ and the selected method/algorithm, the corresponding values of U_{nrep} matrix are substituted with the synthetic one. Replacements are made independently m times to generate m different partially synthetic data sets:

$$SD_i = (U^i_{syn}, U_{nrep}),\ 1 \le i \le m,$$

where U^i_{syn} - a matrix of imputed (replaced) values of i-th synthetic data set. The values in U_{rep} are the same in all synthetic data sets, $SD_i, 1 \le i \le m$.

Thus, the generated partially synthetic data sets $D_{syn} = \{SD_1, SD_2, ..., SD_m\}$ are the information that are provided to the corresponding organizations and the public.

2.2 Analysis of Synthetic Data

Based on the released synthetic data sets $D_{syn} = \{SD_1, SD_2, ..., SD_m\}$, the corresponding organization, in other words the *analyst*, makes inferences about some population quantity $Q = Q(Y)$ (for example, Q can be the average of interest or the coefficients in a regression model). In each synthetic data set SD_i $(i = 1,2,...,m)$, the analyst estimates Q, by some value q_i, and estimates the variance of q_i with some estimator v_i. It is assumed that the analyst determines the q_i and v_i as if SD_i was in fact collected data from a random sample of U. Such technique is usual in area of missing value statistics which needs to develop approaches of unbiased data generation [3].

The approach used in this article is to consider $(q_i, v_i), i = 1,2,...,m$ in similar to [3] as a sufficient characterization of the synthetic databases D_{syn}, and construct an approximate posterior distribution of Q given D_{syn}, $\Pr(Q \mid D_{syn})$, in analogy with the theory of multiple imputation for missing data [5]. In that case the analyst can obtain valid inferences for Q by combining the results of q_i and v_i $(i = 1,2,...,m)$. The following quantities are needed for inferences (see [3]):

$$\overline{q}_m = \frac{1}{m}\sum_{i=1}^{m} q_i,$$

$$b_m = \frac{1}{m-1}\sum_{i=1}^{m}\left(q_i - \overline{q}_m\right)^2,$$

$$\overline{v}_m = \frac{1}{m}\sum_{i=1}^{m} v_i,$$

The analyst then can use $\overline{q}_m$ to estimate the scalar Q and $T_m = (1-\frac{1}{m})b_m - \overline{v}_m$ to estimate the variance of $\overline{q}_m$.

3. Hierarchical Clustering Model for Generation of Partially Synthetic Data

The 2 base approaches we touch in data approximation are as the probabilistic distribution approximation on one hand and the heuristic data extensions on the other. Even when the first one seems more fundamental the heuristic approaches are more interpretable and applicable especially when used by not professionals. Such methods are also fast and can replace successfully the theoretical counterparts in many cases. Having the example of using CART model in synthetic data generation we will try to understand the role and power of hierarchical cluster analysis in the same role. It is more important to understand the inferences of these 2 approaches that complement each other. CART when generated is pruned like the cauterization, and cauterization is evaluated where to stop by the use of additional quality estimates.

3.1 Algorithm for Imputations

Our studies are based on hierarchical agglomerative clustering [14] for generating partially data sets. The principle of synthetic data generation is the same as in the algorithm that considered in [11]. The only difference is that this model is used as a tool for estimating a conditional distribution for sensitive variables in the space where data need to be joined/replaced. That means that the replacement of sensitive variables occurs sequentially (in descending order of number of replacement values). Moreover, for generating new/synthetic values the Bayesian bootstrapping [15] is used. Formally, this process can be described as follows.

Assume that the variables $Y = \{Y_1, Y_2, ..., Y_p\}$ are continuous. Without loss of generality, suppose that from the set of variables $Y = \{Y_1, Y_2, ..., Y_p\}$ the first d is confidential. In addition, the matrix $U_{obs} = (U_{rep}, U_{nrep})$ consists of the first n records of the information units set U.

First, for each variable Y_k $(1 \le k \le d)$, based on the indicator matrix Z, the number of its replacements, $\sum_{i=1}^{n} z_{ik}$ is computed. As a result, $Y_1, Y_2, ..., Y_d$ variables are sorted in descending order of the computed values and denoted as: $Y_{(1)}, Y_{(2)}, ..., Y_{(d)}$.

After that, sequentially confidential variable imputations are processed. Let, $Y_{(k)}$ be the current variable. Firstly, for $Y_{(k)}$ we estimate the conditional distribution in the space where data need to be replaced by using hierarchical agglomerative clustering. Obviously, clustering will be produced based on information units satisfying the following condition:

$$z_{i(k)} = 1, i = 1,...,n\,.$$

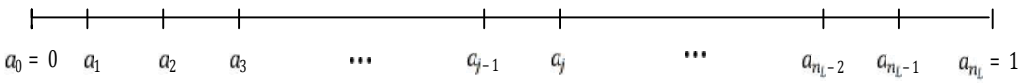
Consider that the quantity of the mentioned units equals to n_k. At the beginning of clustering each information unit is considered as a single cluster: $C_{init} = \{C_1, C_2, ..., C_{n_k}\}$. After that the sequential process of clusters merging begins: each time it brings together a pair of closest clusters, where the distance between them is taken as a distance of their centers, and as an integrated distance measure the Euclidean distance is used:

$$d(C_i, C_r) = \sqrt{\sum_{j=1}^{p} (c_{ij} - c_{rj})^2},$$

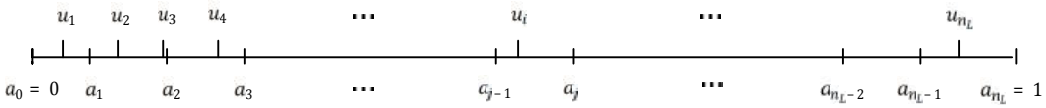
where $ci = (ci1,...,cip)$ and $cr = (cr1,...,crp)$ are centers of clusters C_i, C_r, respectively. To limit the disclosure risk this process continues until there is a possibility of clusters merging in accordance with one of the homogeneity validity measures [16]: *SPR (Semi-partial R-squared),RMSSTD (Root-mean-square standard deviation),* etc. As a result we get the set of current/final clusters $C_{fin} = \{C_1', C_2', ..., C_t'\}$.

Further, in each cluster C_l' $(1 \le l \le t)$ new values for $Y_{(k)}$ are generated by using Bayesian bootstrap procedure. Consider Y^l as a set of values of $Y_{(k)}$ in the corresponding cluster C_l', $Y^l = \{Y_1^l, Y_2^l, ..., Y_{n_l}^l\}$. Bayesian bootstrap draws values based on some donor pool. In this algorithm Y^l is taken for C_l' as a donor pool. Bayesian bootstrap method proceeds as follows:

1. Draw $(n_l - 1)$ uniform random numbers in the range $(0,1)$ and sort these numbers in ascending order: $a_0 = 0, a_1, a_2, ..., a_{(n_l-1)}, a_{n_l} = 1$.



2. Draw n_l uniform random numbers in range $(0,1]$: $u_1, u_2, ..., u_{(n_l-1)}, u_{n_l}$



3. For each u_i $(1 \le i \le n_l)$ determined interval in which it is contained: $u_i \in (a_{j-1}, a_j]$ and replace Y_i^l value by Y_j^l.

Note that some information units involved in clustering, may include new / synthetic values of $Y_{(1)}, Y_{(2)}, ..., Y_{(k-1)}$ variables. In order to maintain imputation consistency for all possible combinations of $Y_{(k)}$ value and $Y_{(1)}, Y_{(2)}, ..., Y_{(k-1)}$ new values corresponding information units are searched, after what $Y_{(k)}$ imputations are done among them.

Thus, as a result of sequential imputations of confidential variables values generated set of partially synthetic data. The whole process is repeated independently m times and generated synthetic databases D_{syn} are provided to the public.

3.2 Simulation Studies

For the above mentioned method our simulation studies are based on public release data from 2011 R.A. Household's Integrated Living Conditions Survey which consist of $N = 7872$ records. Of the entire set of attributes/variables that characterize these units, we are interested in only six of them. The interest variables descriptions are presented in Table 1.

Table 1: Description of variables used in empirical studies

Variable	*Type*	*Range*	*Notes*
Monitory income	Numeric(18,10)	0 – 3512850	
Food purchase	Numeric(18,10)	0 – 555600	
Food consume	Numeric(18,10)	0 - 193191.8083296074	
Nonfood purchase	Numeric(18,10)	0 – 965100	
Expenditures	Numeric(18,10)	6256.3126710940 - 4672865.1342223603	13.643% of house holders have total income more than 225000 AMD
Total income	Numeric(18,10)	0 - 3529697.3182837632	13.795% of house holders have expenditures more than 175000 AMD

Next, we assume that *total income* and *expenditures* are sensitive variables and set threshold conditions for *total income* and *expenditures* are as follows; $total\ income > 225000$ and $expenditures > 175000$.

In empirical studies each observed data set D consists of $n = 315$ randomly sampled households from the 7872 households. As a result of simulation $m = 5$ partially synthetic data sets $SD_1, SD_2, ..., SD_m$ are constructed for each D. Each $SD_i (i = 1,2,...,m)$ is generated using the algorithm presented earlier. Clustering for each sensitive variable is performed on the basis of the units that satisfy the threshold conditions for that variable. In turn, as a validity measure we use minimal count of units in cluster and minimal count of distinct values of sensitive variable in cluster. In this simulation we require minimum ten units with at least three distinct values in each cluster.

Table2: Simple estimands for sensitive variables.

Estimand	Q	*Avg.* $\overline{q_5}$
% of H.H. with total income > 300000	5.206344	5.3860208
% of H.H with expenditures > 230000	6.031744	6.2291976
Average of total income	132357.290182	132611.3233328
Average of expenditures	109301.86175	109548.668466
Standard deviation of total income	94569.735448	95539.39169016
Standard deviation of expenditures	77458.3434	77376.82509436

Table 3: Regression model. Dependent variable: total income. Independent variables: monitory income, food purchase, expenditures.

Estimand	Q	*Avg.* $\overline{q_5}$
Coefficients		
Constant	11279.3796	14181.77272
Monitory income	0.9662	0.90748
Food purchase	-0.2676	-0.20408
Expenditures	0.1652	0.16512
R	0.9844	0.93436

Table 4: Regression model. Dependent variable: total income. Independent variables: monitory income, expenditures.

Estimand	Q	*Avg.* $\overline{q_5}$
Coefficients		
Constant	8868.7488	11893.24448
monitory income	0.9638	0.89952
expenditures	0.072	0.09412
R	0.9818	0.93268

Table 5: Regression model. Dependent variable: expenditures. Independent variables: total income, food purchased, food consumed, nonfood purchased.

Estimand	Q	*Avg.* $\overline{q}_5$
Coefficients		
Constant	-1664.4878	-825.07984
total income	0.026	0.06612
food purchased	1.017	1.00768
food consumed	0.9924	0.90364
food nonpurchased	1.0886	0.906
R	0.9842	0.90692

Table 6: Regression model. Dependent variable: expenditures. Independent variables: food purchased, food consumed, food nonpurchased.

Estimand	Q	*Avg.* $\overline{q}_5$
Coefficients		
Constant	-168.794	3725.97836
food purchased	1.0334	1.05348
food consumed	1.0162	1.204032
food nonpurchased	1.1082	0.95308
R	0.9838	0.90272

Tables 2 – 6 summarize the results of simulation for a variety of estimands. Inferences are made using the methods presented in Section 2.2. For simple estimands (table 2) the averages of synthetic point estimates are close to their corresponding *Q*. The average of parameter R^2 for each regression models for synthetic data sets (tables 3-6) are greater 0.900 which indicates that these models are worth considering. In addition, the averages of regression coefficients are close to original values. So, the analyst will make the same inferences as in the case of actual data.

In case for disclosure risk of each sensitive variable Y_{j_l} *(l=1, 2,…, d)*, we assume that the analyst would estimate Y_{j_l} value of i_r's unit by averaging the replaced values

$$\bar{y}_{i_{r'} j_l} = \sum_{k=1}^{m} y_{i_{r'} j_l}^{rep,k}.$$

[1] R shows how much the independent variables explain the dependent variable

To assess that risk we calculate the root mean squared error (*RMSE*) and relative root mean squared error (*RelRMSE*) of this estimator for each information unit:

$$RMSE_{ir,jl} = \sqrt{(y_{ir,jl} - \bar{y}_{ir,jl})^2 + \sum_{k=1}^{m} (y_{ir,jl}^{rep,k} - \bar{y}_{ir,jl})^2 / ((m-1)m)},$$

$$\mathrm{Re}lRMSE_{ir,jl} = RMSE_{ir,jl} / y_{ir,jl},$$

For any data set, the distributions of the $RMSE_{ir,jl}$ and $RelRMSE_{ir,jl}$ across all units with replaced values can be examined to ensure sufficient variability in imputations. Table 7 displays averages of these quantities across all simulations. Median of *RelMSEs* is typically around 13.5%, which indicates that imputations for most records have a wide range of uncertainty. In case when data owner requires larger errors in terms of decreasing sensitive variables disclosure risk, stricter validity measure criteria can be used in clustering.

Table 7: Sensitive variables limitation in simulation studies.

Variable	*Min.*	*1st Quartile*	*Median*
RMSE			
Total Income	5097.683	16711.692	35692.008
Expenditures	5994.552	22308.272	42092.148
RelRMSE			
Total Income	0.019	0.062	0.10
Expenditures	0.031	0.092	0.168

4. Conclusion

The simulation results show that the proposed clustering model can be used as an alternative approach in partially synthetic data generation in the similar to the CART way. The only limitation is that the attributes/variables characterizing the information units must be continuous. The foregoing can serve as a reasonable prerequisite for the development of clustering model in order to generate synthetic data sets for mixed information units with continuous and categorical attributes.

References

[1] L. Willenborg and T. de Waal, *Elements of Statistical Disclosure Control*, New York: Springer-Verlag, 2001.

[2] D.B. Rubin, "Discussion: statistical disclosure limitation", *Journal of Official Statistics*, vol. 9, pp. 462–468, 1993.

[3] T. E. Raghunathan, J. P. Reiter and D. B. Rubin, "Multiple imputation for statistical disclosure limitation", *Journal of Official Statistics*, vol. 19, pp. 1–16, 2003.

[4] J. P. Reiter, "Significance tests for multi-component estimands from multiply-imputed, synthetic microdata", *Journal of Statistical Planning and Inference*, vol. 131, pp. 365 – 377, 2005.

[5] D. B. Rubin, *Multiple Imputation for Nonresponse in Surveys*, New York: John Wiley and Sons, 1987.

[6] R.J.A. Little, "Statistical analysis of masked data", *Journal of Official Statistics*, vol. 9, pp. 407–426, 1993.

[7] W. Alvey and B. Jamerson, (eds), *Record Linkage Techniques*, Washington, D.C.: National Academy Press., 1997.

[8] J. M. Abowd and S. D. Woodcock, *Disclosure Limitation in Longitudinal Linked Data. Confidentiality, Disclosure, and Data Access: Theory and Practical Applications for Statistical Agencies*, Amsterdam: North-Holland, 2001

[9] F. Liu and R. J. A. Little, "Selective multiple imputation of keys for statistical disclosure control in microdata", *ASA Proceedings of the Joint Statistical Meetings*, pp. 2133–2138, 2002.

[10] J. Drechsler, *Synthetic Datasets for Statistical Disclosure Control. Theory and Implementation*, Springer, 2011.

[11] J.P. Reiter, "Using CART to generate partially synthetic, public use microdata", *Journal of Official Statistics*, vol. 21, pp. 441-462, 2005.

[12] L. Breiman, J. H. Friedman, R. A. Olshen and C. J. Stone, *Classification and Regression Trees*, Belmont, CA: Wadsworh, Inc. , 1984.

[13] J.P. Reiter, "Inference for partially synthetic, public use microdata sets", *Survey Methodology*, vol. 29, pp. 181–189, 2003.

[14] C.D. Manning, P. Raghavan and H. Schutze, *Introduction to Information Retrieval*, Cambridge University Press, 2008.

[15] D.B. Rubin, "The Bayesian bootstrap", *The Annals of Statistics*, vol. 9, pp. 130–134, 1981.

[16] M. Halkidi, Y. Batistakis and M. Vazirgiannis, "Clustering validity checking methods: Part II", *ACM New York, NY, USA*, vol. 31, pp. 19-27, 2002.

Submitted 05.09.2013, accepted 18.10.2013.

L. Aslanyan ,V. Topchyan

Մասնակի սինթետիկ տվյալների գեներացիայի հիերարխիկ կլաստերային վերլուծություն

Լ. Ասլանյան և Վ. Թոփչյան

Ամփոփում

Կոնֆիդենցիալ տեղեկությունների բացահայտման ռիսկի նվազեցումը այսօրվա դրությամբ հանդիսանում է վիճակագրական ընկերությունների հիմնական խնդիրներից մեկը: Այդ խնդրի լուծման համար կիրառվող ամենահայտնի մեթոդներից մեկն է այսպես կոչված սինթետիկ տվյալների բազմությունների մշակման և տրամադրման մեթոդը: Այլ կերպ ասած, կոնֆիդենցիալ փոփոխականների արժեքները փոխարինվում են նոր սինթետիկ արժեքներով: Տվյալ հոդվածում ներկայացված է մասնակի սինթետիկ տվյալների ստեղծման/գեներացիայի մի մեթոդ, որի հիմքում ընկած է հիերարխիկ կլաստերային վերլուծությունը:

Иерархический кластерный анализ для генерации частично синтетических данных

Л. Асланян и В. Топчян

Аннотация

Ограничение риска раскрытия конфиденциальной информации на сегодняшний день является одной из основных задач статистических агентств. Одним из широко применяемых подходов является предоставление синтетических множеств данных. Иными словами, благодаря множественным замещениям значения конфиденциальных переменных исходных данных заменяются новыми синтетическими значениями. В данной статье представлен метод генерации частично синтетических данных на основе иерархического кластерного анализа.