UDC 004.032.26

Yevgeniy V. Bodyanskiy, Alina Yu. Shafronenko

TABLES OF DATA WITH GAPS RESTORATION USING MULTIVARIATE FUZZY EXTRAPOLATION

Annotation. *The problem of the missing values in the data tables filling by using the method of multivariate fuzzy extrapolation is proposed.*

Key words. *Data Mining, data with gaps, multivariable fuzzy extrapolation, neural networks.*

INTRODUCTION

In many Data Mining problems, associated with the processing of information presented as the table “object – property” data may contain missing values (gaps), information in which is lost. The problem of the missing values restoring has received sufficient attention [1-3], in this case as the most effective now are neural networks [4-8]. However, methods of missing values filling based on the restoring of hidden dependencies which are data in the table realization, i.e. explicitly or implicitly in the process of filling the "gaps" is synthesized mathematical model of the phenomena that are described by the table, i.e. solve the identification problem [9, 10].

In practice, the situations often occurs when the dimension of the feature vectors in the table and the number of observations are the same order, i.e., amount of date for mathematical model synthesis is not enough. In this case, the methods of space extrapolation [11] that allow to construct estimates of vector field values using little sets of individual observations can be successfully used.

As one of the most effective approaches based on space extrapolation method the multivariable linear extrapolation (MLE) [11, 12] can be used for recovery of linear functions in the case of insufficient number of observations. MLE can be used for the restoration of non-linear dependencies too, in this case the nonlinear function is determined for not all available observations, but using nearest situation in sense of adopted metrics.

The main disadvantage of MLE is its numerical complexity because the method is based on solving the optimization problem associated with finding the orthogonal projection onto the set of vectors undistorted by missing values. In this case it is necessary to solve the pseudoinversion task [13] for high-dimensional matrices. The following limitation of the method is the fact that in the original table “object - property” number of vectors of observations without missing values must exceed the number of “bad” vectors with gaps. If all the vectors of the table contain “missing

values", MLE is no effective.

In this situation, it seems appropriate to develop simple and effective method for recovering of missing values with a large number of gaps, and in this case instead of the traditional metrics if is convenient to use the concept of membership levels adopted in fuzzy systems and neural networks that are now form main direction in Computational Intelligence [14].

1. PROBLEM STATETMENT

Let we have usual table "object - property" that is shown in Table 1

Table 1

	1	...	*p*	...	*j*	...	*n*
1	x_{11}	...	x_{1p}	...	x_{1j}	...	x_{1n}
...	...	...	...	...	...	...	...
i	x_{i1}	...	x_{ip}	...	x_{ij}	...	x_{in}
...	...	...	...	...	...	...	...
k	x_{k1}	...	x_{kp}	...	x_{kj}	...	x_{kn}
...	...	...	...	...	...	...	...
N	x_{N1}	...	x_{Np}	...	x_{Nj}	...	x_{Nn}

that contain information about N – objects each of that is described by (1×n) - row feature vectors $\underline{x}_i = (x_{i1}, \ldots, x_{ip}, \ldots, x_{ij}, \ldots, x_{in})$. Let's assume that N_G rows may have one or more missing values, and $N - N_G$ ones- completed in full. This does not exclude the situation when $N_G = N$ i.e. all vectors contain the missing values, the number of which in each row $n_i < n, i = 1, 2, \ldots, N.$

During processing, the table must be filled in the missing values so that the recovered elements were in some sense the most "similar" or "nearest" to a priori unknown regularities hidden in this table.

2. THE MULTIVARIABLE FUZZY EXTRAPOLATION METHOD

Let's represent Table 1 in the form of $(N \times n)$ - matrix X, in which, in the simplest case are absent one element x_{ip} or more generally $\sum_{i=1}^{N} n_i$ elements. All data are previously centered and standartized by all features, so that all observations belong to the hypercube $[-1, 1]^n$. Therefore, the data form array $\tilde{X} = \{\tilde{x}_1, \ldots, \tilde{x}_k, \ldots, \tilde{x}_N\} \subset R^n$, $\tilde{x}_k = (\tilde{x}_{k1}, \ldots, \tilde{x}_{ki}, \ldots, \tilde{x}_{kn})^T$, $-1 \le \tilde{x}_{ki} \le 1$,

$1 < m < N$, $1 \le q \le m$, $1 \le i \le n$, $1 \le k \le N$. For each row $\underline{x}_i$, containing missing values we have to estimate distances between it and all the other rows using the concept of "partial distance" (PD), adopted in fuzzy clustering [15] and modified as

$$D_P^2(\underline{x}_i, \underline{x}_k) = \frac{n}{n_i + n_k - n_{ik}} \sum_{j=1}^{n} (x_{ij} - x_{kj})^2 \delta_j$$

where

$$\delta_j = \begin{cases} 0, \text{ if } \underline{x}_i, \text{ or } \underline{x}_k \text{ in } j \text{ position contains missing value,} \\ 1 \text{ otherwise,} \end{cases}$$

n_{ik} -total number of missing values in the same position in $\underline{x}_i$ and $\underline{x}_k$.

In this case we have to exclude from consideration $\underline{x}_k$ for which $\sum_{j=1}^{n} \delta_j = 0.$

Let us order further $\tilde{N} \le N - 1$ calculated distances so that

$$0 \le D_P^{2[min]} = D_P^{2[1]} < D_P^{2[2]} < \ldots < D_P^{2[\tilde{N}]} \le 4n$$

(here the index in square brackets indicates the rank) and save for further processing only $\hat{N} \le \tilde{N} \le N - 1$ observations, that satisfy inequality

$$\frac{D_P^{2[l]}}{4n} \le \varepsilon, \quad l = 1, 2, \ldots, \hat{N},$$

where ε - a certain threshold $(0 < \varepsilon < 1)$.

Using the concept of memberships, adopted in the standard fuzzy c-means method [16],let's calculate the membership level $\underline{x}_i$ to $\hat{N}$ still under consideration vectors $\underline{x}_l$ in the form

$$U_l(i) = \frac{D_P^{-2[l]}}{\sum_{q=1}^{\hat{N}} D_P^{-2[q]}}, \quad l = 1, 2, \ldots, \hat{N},$$

while if $D_P^{2[1]} = 0$, we automatically assume that $U_1(i) = 1$.

Thus, each vector $\underline{x}_i$ is approximated by the expression

$$\hat{\underline{x}}_i = \sum_{l=1}^{\hat{N}} U_l(i) \underline{x}_l. \qquad (1)$$

Let's note too that MLE is used as an approximation of type (1), but instead of memberships levels $U_l(i)$ the weights obtained by solving the optimization problem, which is not always solvable are used.

And finally, the last stage - to fill missing values. It is easy to see that the estimate of missing element $\underline{x}_{ip}$ can be written as

$$\hat{\underline{x}}_{ip} = \sum_{l=1}^{\hat{N}} U_l(i)\underline{x}_{lp}.$$

The proposed method can be conveniently represented in the form shown in Figure 1.

3. EXPERIMENTALRESEARCH

The provided results are presented on the fig.2. On this picture shows a database that contains data connected with the X-ray plant. Real data has been corrupted and restored by the proposed method based on multivariate fuzzy extrapolation.

To estimate the quality of the algorithm we used the mean absolute percentage error (MAPE). When estimating the quality of the recovered data x-ray plant MAPE does not exceed 15 percent.

The problem of restoration of distorted data provided by the x-ray plant using the proposed method, making it possible speed up recovery hardware that is out of order.

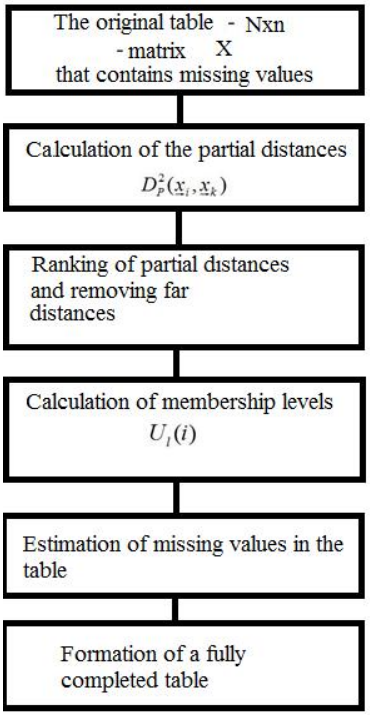

Fig. 1 - The method of multivariate fuzzy extrapolation in the task of restoring missing values

 ISSN 1562-9945

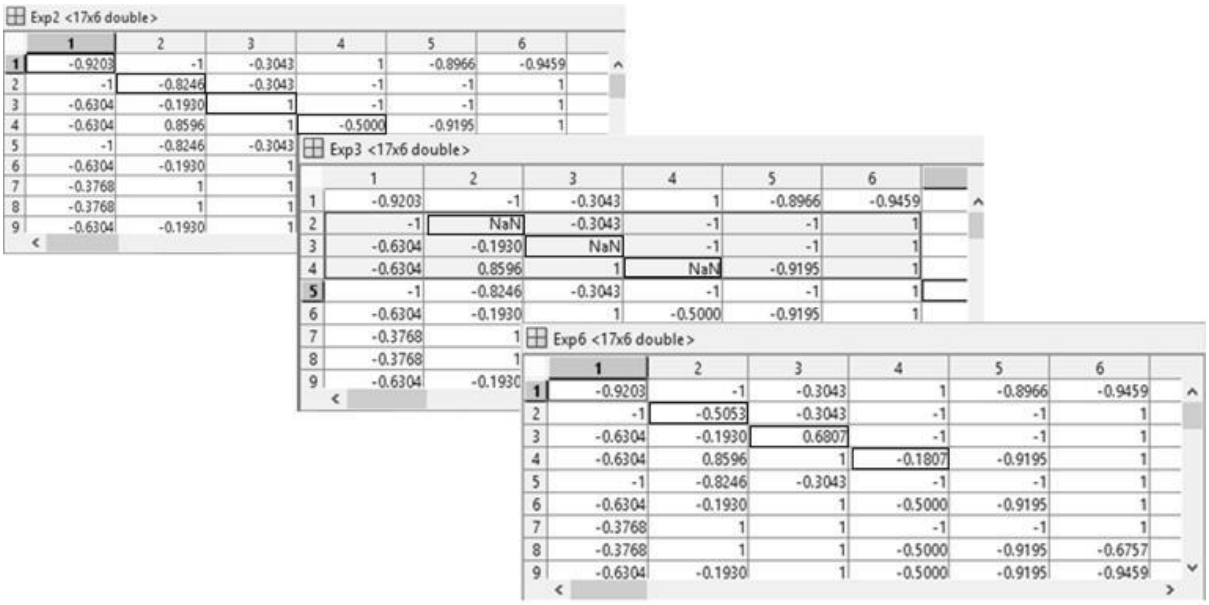

Fig.2 – Results of experiments

CONCLUSION

The problem of the missing values in the data tables filling by using the method of multivariate fuzzy extrapolation is proposed. The method has clear physical sense, derived from the theory of fuzzy systems, and characterized by computing simplicity and high speed data processing.

REFERENCES

1. Загоруйко Н.Г. Эмпирические предсказания – Новосибирск: Наука, 1979. – 120с.
2. Han J., Kamber M. Data Mining: Concepts and Techniques. – Amsterdam: Morgan Kaufman Publ., 2006. – 743p.
3. Gorban A., Kegl B., Wunsch B., Zinovyev A.(Eds.) Principal Manifolds for Data Visualization and Dimension Reduction. Lecture Notes in Computational Science and Engineering, Vol. 58. – Berlin– Heidelberg – New York: Springer, 2007. – 330 p.
4. Bishop C.M. Neural Networks for Pattern Recognition. – Oxford: Clarendon Press, 1995. – 482p.
5. Gorban A.N., Rossiev A.A., Wunsch II D.C. Neural network modeling of data with gaps// Radioelectronics. Informatics. Control. – 2000. – №1 (3). – С. 47 – 55.
6. Tkacz M. Artificial neural networks in incomplete data sets processing // In: Eds. Klopotek M.A., Wierzchon S.T., Trojanowski K. Intelligent Information

Processing and Web Mining. – Berlin – Heidelberg: Springer – Verlag, 2005. – P.577 – 583.

7. Marwala T. Computational Intelligence for Missing Data Imputation, Estimation, and Management: Knowledge Optimization Techniques. – Hershey – New York: Information Science Reference, 2009. – 303p.
8. ПліссІ.П., ШевяковаА.Ю., ШевяковаЮ.Ю. Нейромережеве відновлення пропусків у таблицях даних//Наукові праці. – Миколаїв, Вид-во ЧДУ ім. Петра Могили. – 2011. – Вип. 148. – Т.160. Комп'ютерні технології. – С. 59-61.(in Ukrainian)
9. Ljung L. SystemIdentification: TheoryforUser. – Sweden: Prentice-Hall, Inc., 1987. – 432c.
10. Nelles O. NonlinearSystemIdentification. – Berlin: Springer, 2001. – 785p.
11. Растригин Л.А. Адаптация сложных систем. – Рига: Знатне, 1981. – 375 с.
12. Растригин Л.А., Пономарев Ю.П. Экстраполяционные методы проектирования и управления. – Москва: Машиностроение, 1986. – 120с.
13. Albert A. Regression, Pseudoinversion, and Recursive Estimation [Russian translation]. -Moscow, 1977.
14. Rutkowski L. Computational Intelligence. Methods and Techniques. – Berlin Heidelberg: Springer-Verlag, 2008. – 514p.
15. R.J. Hathaway, J.C Bezdek. Fuzzy c-means clustering of incomplete data. IEEE Trans on Systems, Man, and Cybernetics, №5, 31, 2001, P. 735-744.
16. J.C. Bezdek. Pattern Recognition with Fuzzy Objective Function Algorithms. Plenum Press, New York, 1981.
17. Michel A.N., Farrel J.A. Associative memories via artificial neural networks//IEEE Control Systems Magazine. – 1990. – 10(3). – P.6-17.
18. Hassoun M.H. Fundamentals of Artificial Neural Networks. – Cambrige, MA: MIT Press, 1995. – 452p.
19. Rojas R. Neural Networks. A Systematic Introduction. – Berlin: Springer-Verlag, 1996. – 502p.
20. Hagan M.T., Demut H.B., Beale M. Neural Networks Design. – Boston: PWS Publishing Company, 1996.-729p.
21. Hassoun M.H., Watta P.B. Assosiative memory networks/ In "Handbook of Neural Computation". – Oxford: IOP Publishing Ltd. and Oxford University Press, 1997. – p.3:14.

22. Haykin S. Neural Networks. A Comprehensive Foundation. – N.Y.: Prentice Hall, Inc., 1999. – 842p.
23. Bodyanskiy Ye., Teslenko N. Autoassociative memory evolving system based on fuzzy basis functions// Sci. J. of Riga Techn. Univ. – Comp. Sci. Inf. Technology and Management Sci. – 2010. – 44. – P.9-14.