UDC 004.6

N.V. Kuznietsova, P.I. Bidyuk

National Technical University of Ukraine “KPI”, Kyiv, Ukraine

BUSINESS INTELLIGENCE TECHNIQUES FOR MISSING DATA IMPUTATION

Background. Properly constructed decision support systems (DSS) for modelling and forecasting behaviour of dynamic systems provide a possibility for taking into consideration uncertainties of probabilistic, statistical and structural types what results in higher quality of developed models and estimated forecasts.

Objective. To consider general reasons for loosing (missing) data in statistical data analysis; to provide categorization of missing data into several groups, and identify the reasons for missing measurements; to provide stepwise system methodology for uncertainty analysis and selection of data imputation techniques; to give an insight into some popular missing values imputation techniques regarding their possible applications.

Methods. To solve the problems mentioned the following methods have been used: data categorization approach from business or practical point of view that is necessary for discovering the reasons for availability of systemic and/or random missing values; the modern systemic methodology was hired for analysis of uncertainty causes and missing values imputation; the decision tree based imputation procedures; EM algorithm and regression model approach to forecasting missing data using forecasting functions.

Results. The main results of the study are in categorization of the missing data into groups; development of systemic methodology for analysis of uncertainty causes and missing values imputation; providing an analysis for possibilities of missing values imputation with decision trees, EM algorithm and regression models; development of multistep forecasting functions on the basis of autoregression models; illustration of application of some selected perspective methods for missing data imputation.

Conclusions. We proposed the six steps system methodology for data imputation which stresses that selection of correct method for imputation is tightly connected with the step-by-step analysis of the gaps causes and finding an appropriate technique for their imputation. The results of imputation sometimes are rather far from the existing data and should be smoothed or even broken from the sample due to their incorrectness. For such cases it should be proposed a new probabilistic-regression method which allows define parameters of the probability interval for the regression aiming missing data imputation. A series of computing experiments performed with EM algorithm, forecast regression based imputation technique and some other approaches shows that it is possible to reach high quality results regarding correct processing of data with missing values.

Keywords: uncertainties in data processing; imputation of missing data; systemic approach; decision support system.

Introduction

Data quality is the main factor that influences success of model building, estimating forecasts and the value of dynamic system state prediction. Very often researchers are facing the problem of data incompleteness, inaccuracy, influence of noise and random external disturbances. The decision regarding removing such data from a sample could be implemented only for the huge amounts of data in a sample that don’t contain key data for forecasting. In practice the problem of missing data appears first of all in those areas where each raw in data sample, even containing missing values, is a key data and should be involved in decision making process.

Many special methods for imputation of missing data have been developed within the last forty years. The use of specific methods and techniques depends on particular problem area where missing data happen [1–3]. The missing values can result from data collection errors, incomplete customer responses, actual measurement system failures, or from a revision of the data collection scope over time, such as tracking new variables that were not included in the previous data collection schema. It is known from the previous studies that hiring of appropriate missing data imputation schemes may lead to high quality final results such as forecasts estimates and the decisions based on the forecasts.

This study is directed towards categorization of missing data and development of a stepwise methodology for uncertainty analysis and selection of data imputation techniques especially for the cases when data are given in the form of time series.

Problem statement

The goals of the study are as follows: to consider the general reasons for loosing (missing) data in statistical data analysis; to provide categorization of missing data into several groups, identify the reasons for missing measurements; to develop stepwise system methodology for uncertainty analysis and selection of data imputation techniques; to give an insight into some popular missing values imputation techniques regarding their possible applications.

General considerations

Data incompleteness is rather widespread fact in data processing and decision making processes. The reason for it could be explained by the operator failures who gathered the data or by the circumstances such as impossibility to get some definite kind of data, time delay or missing features for specified data. Some part of missing data could be adjusted from another data sources but of course it requires some time and often substantial efforts of a researcher. The cost of the efforts could be even more expensive than usage of specified techniques for missing data evaluation and imputation.

For example, a bank should contact every customer with a missing value; or contact real estate representative and ask for necessary (missing) characteristics [3]. It could be easier to increase the percentage of such missing data for example by proposing special benefits for real estate holder (say, +1 % to real estate holders or special 20 % discounts for real estate insurance by insurance company partner). Even if percentage of missing data is not large in data sample but this kind of data could be really important in the future and couldn't be excluded from training sample while incompleteness of data could be normal fact for new clients. Unfortunately not all developed techniques can model missing data effectively. Sometimes the results of modeling received by different data mining methods quite differ from the real data. In this case the most direct approach is in using of mean or median for missing data filling in the data sample.

Nevertheless, the use of such methods injure sample and smooth it that could be really useless for defining splashes and critical points. Choosing the "best" missing value replacement technique inherently requires from a researcher to make assumptions about the true (missing) data. For example, researchers often replace a missing value with the mean of a variable [4]. This approach assumes that the variable's data distribution follows a normal population response. Replacing missing values with the mean, median or another measure of a central tendency is simple, but it can greatly affect an actual variable's sample distribution [5]. The results of missing data filling for different spheres show that new methods should be developed for imputation which would allow achieving more precise results regarding correct approximation of actual data.

Missing data: brief overview

Missing data can be categorized into the following groups [2, 6]:

- The feature is not applicable for some subjects. From a business or practical point of view this means that for certain subject(s) of analysis no new value can be set for a certain feature. I. e., the respective fact is not applicable for this special record.
- The feature is applicable, but it has not been (or could not be) retrieved. Here the data would be available; however, they have not been provided in the data collection or business process. The reasons for this could be like that:

– The information was not provided. For example, a person did not provide the number of people in his household.

– Certain data have been provided; however, they have not been processed further on, for example, in data collection or data entry. If the existence of the true value has high priority, additional efforts could be invested to receive this value (for example, getting a new data entry, interviewing the customer again in a market survey, or inspecting the paper forms).

– Data have been entered into the system, but the value or the record has been deleted during data transfer or data management. If the original data still exist, data transfer can be executed again, and backup copies of the data could be retrieved and processed with revised programs.

- For various reasons, some data values have been entered incorrectly and could be immediately recognized as senseless or obviously wrong. This is not only a correctness problem, but it is also a completeness problem because these values often need to be set to missing.

In many cases, completeness of the data can be derived from the fact that values exist for a certain column in a table. However, to judge completeness only from the fact that values exist or not may lead in the wrong direction: if a non-missing value exists in a table, it does not necessarily mean that a valid value exists for a certain observation. In many cases, missing values are entered as values like 0 or 999, which from a technical point of view is non-missing but not from a business point of view. Also, for categorical variables, a missing value may be coded with a separate category.

In contrast, the fact that a variable has missing values does not mean that no information is available in this case:

- The data representation of answers to multiple choice questions in a survey is often only entered as one value for those categories that were checked. The non-checked items are left as a missing value in the data. However, they truly mean "not checked" or "does not apply". In this case, no value other than replacement value of zero (0) makes sense, and there is no need for a complicated imputation method to correct the error.
- For a customer who has no calls to the call center, a variable holding the number of calls may show a missing value, indicating that no call has been made. This value should then be replaced by zero.
- In the context of transactional or time series data, frequently the count for intervals or categories with no observations is represented as a missing value, which, however, should be interpreted as a zero value.
- Also a variable may be missing for some observations, but its values can be calculated from other variables of the same subject under investigation.

If observations contain a missing value, then by default that observation is not used for modeling by techniques such as neural network, Bayesian network or regression [7]. However, rejecting all incomplete observations may ignore useful or important information which is still contained in the non-missing variables. Rejecting all incomplete observations may also bias the sample, since observations that missing values may have other things in common as well.

Missing values can be categorized into two categories: systematic and random. The differentiation between the two groups is important in order to decide the impact of that value on the analysis results and on the options in the treatment of missing values [1].

Random missing values. Random missing values are defined by the fact that each observation in the data has the same probability of having a missing value for a certain variable. The fact that a value is missing does not depend on other variables in the data mart or on other causalities. If all data for observations were available initially, random missing values could be generated by deciding for each observation whether the value should be set to missing. The process of handling the data applies to numeric and categorical data equally. In descriptive statistics this usually leads to separate specification of the number of missing values and the calculation of respective statistics from the non-missing values.

The decision about randomness in the occurrence of a missing value is, however, usually not performed on very strict criteria. Any missing value can be considered to have a causal background. Thus it can be a philosophic discussion to decide whether there is any case where a missing value occurs purely at random, or based on a possibly hidden systematic basis.

Random missing values decrease the amount of information in the analysis database. Valid observations are available for fewer observations. This means that even if effort has been made to include many observations in the analysis or analysis database, a subset of them cannot be used.

Missing values that are truly random do not damage the data, their distribution, their relationship to other variables, and the inferences that can be drawn from analyses because such missing values do not introduce bias (the picture that is seen is incomplete, but it is not wrong). Still, truly random missing values affect the precision of the analysis, though in large sample sizes the effect is quite small.

In analytical methods like regression analysis, for example, observations with missing values cannot be used in the analysis because no value can be inserted into the regression equation. In multiple regression methods, where more than one variable is used, a missing value for one variable, however, means that all the observations cannot be used in the analysis, and the existing information for other variables is ignored as well. In these cases it is better to decide how missing values could be replaced with a most likely value and to use the observation in the analysis. Decision trees are not as vulnerable to missing values because they treat a missing value as a separate category [5, 7].

In the case of random missing values the missing values are either considered as a separate category with the assumption that the statistics shown for the non-missing values are representative or imputation logic can be found on how to derive a most representative replacement value.

The second example happens in data mining analyses or selected types of statistical analysis. In some disciplines like clinical research the imputation of missing values is not performed at all because the analysis results must be based only on real data. A good point with random missing values is that not only does the existing data represent a true picture for the analysis, but also that imputation methods can be used to impute the observations with missing values.

Say SAS platform offers a range of methods to impute missing values, including methods for time series and methods for one-row-per-subject data marts. Some of these methods impute a static value for all observations; other methods impute an individual most likely value [1]. The challenge in this case is to define the percentage of missing values that is still acceptable for imputation and allows one to get meaningful results or representative replacement values.

Systematic missing values. Systematic missing values are more difficult to deal with because the non-missing values cannot be considered as a representative sample of the population. Thus, more elaborate replacement methods and strategies need to be applied.

The crucial input to these methods is the knowledge about the origin of the data, the business process, the data collection process, and the reason that they are missing, to name a few. This knowledge is needed to formulate the imputation logic for the missing data.

For some problems where missing data have systemic character it is easy to find some period or some law for locating missing data. Even not substantial information about boundaries or statistics of the missing data distribution gives the possibility for correct imputation. Treating missing values as random cases and replacing them with average values will bias the data and the resulting analysis. Thus, more thought is to be put into handling these data in order to overcome this data quality issue of missing information.

It could be useful to compare the distribution of the existing data with the assumed distribution. This assumed distribution can be based on business considerations, on market research or even general assumption. A system of rules for the replacement of the missing values can then be built so that replacing the missing values would result in the assumed distribution.

For coping with missing interval variables the following imputation techniques such as Andrew's Wave, Default Constant, Distribution, Huber, Mean, Median, Mid-Minimum Spacing, Midrange, None, Tree, Tree Surrogate, Tukey's Biweight are provided in SAS [1, 3, 7]. The wide variety of methods is specified due to the infinite number of possible states to the interval variable by definition. For example, for class variables are proposed such imputations techniques for missing data as Count, Default Constant, Distribution, None, Tree, Tree Surrogate. Default methods for both types of variables (interval or class variables) allow customize the default imputation statistics by specifying own replacement values for missing and non-missing data. Missing values for the training, validation, test, and score data sets are replaced using imputation statistics that are calculated from the active training predecessor data set. The use of the Decision Tree Node allows defining either decision alternative or surrogate rules in order to group missing values into a special category. One can also use the Cluster node to replace missing values.

To generalize overviewed standard approaches to different types of missing data we propose an appropriate methodology for deep analysis of missing data causes and choosing the best methods for missing data imputation.

Systemic methodology for analysis of uncertainty causes' and missing values imputation

Step 1. Facing with missing data an analysis of a sample classification of missing data should be done. It could be categorical or numerical missing value(s). In a case of categorical variables one of the possible methods for missing values processing is moving them into separate category. For numerical missing values and categorical missing values which still require additional analysis go to step 2.

Step 2. A profound analysis of missing data causes and data sources spaces: for evaluation of missing values first of all it is necessary to study the possibility in general and evaluate needed efforts to restore the missing values. If there is no possible reasons for clarifying missing values the root causes for the missing data should be found and fixed. If the cause for missing data is the fact that such data was not collected before or not collected now, it means that there was some change in data gathering process and such information about changes is fixed in archives for defined products with the date when these changes occurred. If there is no information about such changes, it is obvious that it is necessary to clarify the nature of missing values: a systematic (for example, in some region the staff systematically does not fix some characteristics about clients) or random. If this information is not related to the feature of the system or staff, or some certain failures in the software, it is necessary to determine if the missing data are a real threat.

Step 3. The gaps related to the systemic missing data due to the staff or system failures. The causes of that could be as follows: 1) an accidental omission of information (for example, it is not compulsory to fill customer or client does not contain identification number because of religious beliefs), 2) the client does not want to provide certain information; 3) client makes mistakes in filling out some information and there is a real opportunity to restore (deliberate distortion of information). The last two reasons are quite significant in the analysis of admission and can be used as a the basis for attributing these clients to unreliable and their refusal to provide certain services or goods. Therefore such cases should include a separate category in the next step.

Step 4. In depth analysis of the causes of missing values and determining the worst case (reasons 2 and 3 in Step 3) with replacing the missing value by smallest/biggest value that is the worst option for this variable.

Step 5. The use of special tools and techniques to restore missing values. For different intelligent platforms there are many methods for missing data imputation. Algorithms like Zet, Zetbraid etc., that set the omitted value by finding the most similar characteristics to other indicators in the data set or replace an average value for this characteristic [5].

Step 6. Using the values recovered for further modeling.

Now discuss the tree imputation method as one of the appropriate methods for missing data imputation which is widely used in such platforms like SAS and SPSS.

Tree Imputation

Specifying some parameters for tree imputation analytics could customize it to specific data set and decision of specified task. Variety of the decision tree features, and rules for splitting allows constructing different models for the same data sets and achieve the best results for imputation of missing values by more similar data [1, 3].

◊ **Leaf Size** – specify minimum number of training observations that are allowed in a leaf node. Permissible values are integers greater than or equal to 1. The default setting is 5.

◊ **Maximum Branch** – specify the maximum number of branches that you want a splitting rule to produce. Permissible values for the Maximum Branch property are integers between 2 and 100. The minimum value of 2 results in binary trees. The default value for the Maximum Branch property is 2.

◊ **Maximum Depth** – specify the maximum number of generations of nodes that you want to allow in your decision tree. An original node is the root node. Children of the root node are the first generation. Permissible values are integers between 1 and 100. The default number of generations for the Maximum Depth property is 6.

◊ **Minimum Categorical Size** – specify the minimum number of training observations that a categorical value must have before the category can be used in a split search. Permissible values are integers greater than or equal to 1. The default value for the minimum categorical size property is 5.

◊ **Number of Rules** – specify the number of splitting rules that you want to save with each node. The tree uses only one rule. The remaining rules are saved for comparison. Permissible values are integers greater than or equal to 1. The default value for the Number of Rules property is 5.

◊ **Number of Surrogate Rules** – specify the maximum number of surrogate rules that imputation seeks in each non-leaf node. The first surrogate rule is used when the main splitting rule relies on an input whose value is missing. Permissible values are nonnegative integers. The default value for the Number of Surrogate Rules property is 2.

◊ **Split Size** – specify the smallest number of training observations that a node must have before it is eligible to be split. The Split Size property uses a default value of (2×Leaf Size), unless you specify an integer value that is greater than the calculated default value. If no value is specified for the Leaf Size property, then the default Split Size value is calculated as: [2×Min(5000,Max(5,N/1000))]. Permissible values for the Split Size property are integers between 2 and 32767 [1].

EM algorithm

An expectation maximization algorithm (EM algorithm) is widely used in mathematical and applied statistics, optimization theory and its multiple applications for computing unknown model parameters, imputing lost measurements, finding the minima and maxima values for various functions etc. One of its applications is directed towards maximum likelihood estimation of unknown model parameters for probabilistic models in the cases when some variables cannot be measured directly.

The algorithm is functioning iteratively in two steps [8]. At the E-step (expectation step) an expected value of likelihood function is computed using current approximation of non-measurable variables. The M-step is used for computing the model parameter estimates that maximize the expected likelihood generated at the E-step. The EM algorithm is also often used for data clustering, machine learning, in computer vision systems and natural language processing (it is known as a special case of Baum-Welch algorithm). Due to the possibility of functioning in the conditions of lost data the EM algorithm is very useful instrument for portfolio risks estimation in analysis of financial data. With respect to the missing data problem EM algorithm is applied in two frequent cases: (1) when the measurements are missing due to possible problems with organization of observation process; (2) when likelihood function optimization is analytically impossible but computations can be correctly simplified by assuming existence of additional missing or hidden parameters.

Consider some basics of this approach. Denote as $\mathbf{X}$ matrix of incomplete data, and complete data set as $\mathbf{Z} = [\mathbf{X}, \mathbf{Y}]$ that could be described by joint probability distribution (density) function as follows:

$$p(\mathbf{Z} \mid \theta) = p(\mathbf{X}, \mathbf{Y} \mid \theta) = p(\mathbf{Y} \mid \mathbf{X}, \theta) p(\mathbf{X} \mid \theta),$$

where θ is parameter vector for specific distribution. Thus, we defined joint density function (distribution) between the missing $\mathbf{Y}$ and available measurements $\mathbf{X}$ with corresponding likelihood function

$$L(\theta \mid \mathbf{Z}) = L(\theta \mid \mathbf{X}, \mathbf{Y}) = p(\mathbf{X}, \mathbf{Y} \mid \theta).$$

This function is actually random variable due to the fact that missing data $\mathbf{Y}$ is unknown (random); $\mathbf{X}, \theta$ are considered as known constants (values) for current computational step. The EM algorithm can be specified in two steps as given below [8].

Step 1 (expectation). The EM algorithm has to compute expected value of the complete data log-likelihood function $\mathrm{Log}[p(\mathbf{X}, \mathbf{Y} \mid \theta)]$ with respect to the unknown measurements $\mathbf{Y}$ given observed data $\mathbf{X}$ and current parameter estimates θ_k. Actually at iteration k we will be able to use the parameter estimates taken from the previous iteration θ_{k-1}, i. e. the following log-likelihood is optimized:

$$L_1(\theta \mid \theta_{k-1}) = E[\mathrm{Log}(p(\mathbf{X}, \mathbf{Y} \mid \theta) \mid \mathbf{X}, \theta_{k-1})], \quad (1)$$

where θ is optimal parameter vector that we are looking for. This expression is a conditional expectation with the condition: $(\mathbf{X}, \theta_{k-1})$. Using the marginal distribution for missing values $f(\mathbf{Y} \mid \mathbf{X}, \theta_{k-1})$, the right hand side of (1) can be written in the form:

$$E[\mathrm{Log}(p(\mathbf{X}, \mathbf{Y} \mid \theta) \mid \mathbf{X}, \theta_{k-1})]$$
$$= \int_{y \in \Omega} \mathrm{Log}(p(\mathbf{X}, y \mid \theta) f(y \mid \mathbf{X}, \theta_{k-1})) dy, \quad (2)$$

where Ω is the space of values that variable y could take on. Very often the density that is actually needed is defined as follows:

$$f(\mathbf{y}, \mathbf{X} \mid \theta_{k-1}) = f(\mathbf{y} \mid \mathbf{X}, \theta_{k-1}) f(\mathbf{X} \mid \theta_{k-1}).$$

Practically useful for computations at the first step is the following deterministic expression that can be maximized with respect to θ:

$$q(\theta) = E_{\mathbf{Y}}[\varphi(\theta, Y)] = \int_y \varphi(\theta, Y) f_Y(y) dy,$$

where $\varphi(\theta, \mathbf{Y})$ is a function with θ constant and $\mathbf{Y}$ is random variable governed by its distribution $f_{\mathbf{Y}}(y)$.

Step 2 (maximization). Now it is necessary to maximize the expectation found at the first step:

$$\theta_k = \arg\max_{\theta} L_1(\theta, \theta_{k-1}).$$

The two steps defined are computed iteratively as necessary until some selected stop criterion is fulfilled. At each iteration of executing the algorithm the log-likelihood (1) is increased and EM algorithm should converge to local maximum of this function. The distribution parameters computed this way are to be used for generating extra data from the joint distribution.

Regression approach to missing values imputation

Relatively simple approach to missing values imputation when performing time series analysis is in hiring regression models such as autoregression (AR) and autoregression with moving average (ARMA). It can be applied when we have enough data for constructing AR and/or ARMA models for generating forecast estimates to fill in the missing values. The simplest equation from this subclass of liner models is AR(1):

$$y(k) = a_0 + a_1 y(k-1) + \varepsilon(k), \quad E[\varepsilon(k)] = 0. \quad (3)$$

To find one-step ahead forecast increase discrete time by one:

$$y(k+1) = a_0 + a_1 y(k) + \varepsilon(k+1).$$

Assuming the coefficients a_0, a_1 are known, we can find the forecast as conditional mathematical expectation for the right hand side using all available information about the process AR(1) including the moment k:

$$\hat{y}(k+1,k) = E_k[y(k+1)]$$

$$= E_k[y(k+1)|y(k), y(k-1), ..., \varepsilon(k), \varepsilon(k-1), ...]$$

$$= a_0 + a_1 E_k[y(k)] = a_0 + a_1 y(k),$$

as far as $y(k)$ is known constant (the last measurement available) at the moment k. Using again the same approach find two step-ahead forecast with (3):

$$y(k+2) = a_0 + a_1 y(k+1) + \varepsilon(k+2)\,,$$

$$\hat{y}(k+2,k) = E_k[y(k+2)] = a_0 + a_1 E_k[y(k+1)]$$

$$= a_0 + a_1 E_k[a_0 + a_1 y(k)] = a_0 + a_0 a_1 + a_1^2 y(k).$$

By induction we can write three step-ahead forecast as follows:

$$\hat{y}(k+3,k) = E_k[y(k+3)]$$

$$= a_0 + a_0 a_1 + a_0 a_1^2 + a_1^3 y(k).$$

Thus, s step-ahead forecast can be computed using the following function:

$$\hat{y}(k+s,k) = E_S[y(k+s)] = a_0\left(\sum_{i=0}^{S-1} a_1^i\right) + a_1^S y(k)$$

$$= a_0 \sum_{i=0}^{S-1} a_1^i + a_1^S y(k). \qquad (4)$$

Equation (4) is called forecasting function for an arbitrary number of steps-ahead s. The sequence of forecasts is a convergent process if the condition is fulfilled: $|a_1| < 1$, i. e.

$$\lim_{s\to\infty} E_k[y(k+s)] = \frac{a_0}{1-a_1}\,, \ |a_1| < 1\,, \qquad (5)$$

where a_1 is the denominator of the geometric progression resulting from right hand side (RHS) in (4). Expression (5) shows that for arbitrary stationary AR or ARMA process the sequence of conditional forecast estimates is asymptotically convergent, $s \to \infty$, to unconditional mean. For obvious reason the constant in the RHS of (5) is also called long term forecast.

Extension of the forecasting function to AR(p) process can be written in the form of the following recursion:

$$\hat{y}(k+s,k) = a_0 + \sum_{i=1}^{p} a_i \hat{y}(k+s-i)\,,$$

where $\hat{y}(k+s-i) = E_k[y(k+s-i)]$. If a specific data set is powerful enough to develop several candidate models, then the best model among the candidates should be selected for generating necessary forecasts. Alternatively, we could use the forecasting approach based on combination of forecasts generated by several constructed models for the same time series. Such approach is useful when the variances of separate forecasts errors do not differ substantially, say less than an order.

A number of carried out computing experiments showed that both EM algorithm and regression (forecasting) approach can provide high quality results regarding correct processing of data with missing values. As an imputation example consider a sample consisting of 500 values (stock price data). Initially forecasting model was constructed for the whole data sample. Then 25 data points of this sample were eliminated (from 301 to 325) and several alternative techniques were used for the data imputation. The results achieved are shown in table.

For each data case the best model turned out to be AR(1) with the third order trend (t3). Ac-

Table. Results of data imputation using different techniques

Conditions for model constructing	Best resulting model	Model quality		Forecast quality	
		R^2	DW	$RMSE$	$MAPE$
Complete initial sample	AR(1) + t3	0.993	2.12	25.92	3.19
Lost data is replaced by zeros (real gap)	AR(1) + t3	0.31	0.10	118.14	6.51
Replacing the gaps by sample mean	AR(1) + t3	0.990	1.97	27.97	3.45
Replacing the gaps with forecasts	AR(1) + t3	0.991	2.18	26.78	3.27
Imputation of data using EM algorithm	AR(1) + t3	0.992	2.12	25.97	3.21

cording to the results shown in table the best imputation techniques in this case turned out to be the EM algorithm that provided mean absolute percentage error (MAPE) for one-step prediction of about 3.21 % that is quite comparable with 3.19 % computed for the complete sample. The model adequacy evaluated with the determination coefficient R^2 and the Durbin-Watson statistic DW was practically the same in both cases. The worst result was generated for the case when the missing values were replaced by zeros. In this case MAPE = 6.51 % what is twice as much in comparison to the complete sample. And the determination coefficient decreased by more than three times (from 0.993 to 0.31). As far as it is usually difficult to predict which technique will provide better final result it is preferable to use alternative computational procedures and select the best one for a specific case using known quality statistics.

Conclusions

To decrease the influence of data uncertainty the analysis of available gap causes should be done as well as evaluation of missing values and filling in the gaps using special methods like resampling, EM-algorithms, Hot-deck, Barleta and Zet, etc. The correct evaluation of missing data and using their forecasts for the next modeling is the main task of researchers. Special intelligence platforms such as SAS, SPSS use typical standard algorithms and approaches for gaps analyzing and missing data imputation such as mean, median, constant, decision tree, M-estimators etc. Obviously, they give the same results for imputation with the same parameters when similar methods are used.

Nevertheless, the system methodology proposed in this paper shows that using the correct method for imputation is tightly connected with the step-by-step analysis of the gaps causes and finding the appropriate techniques for their imputation. The results of imputation sometimes are rather far from the existing data and should be smoothed or even broken from initial sample due to their incorrectness. For such cases it should be proposed a new probabilistic-regression method that would allow define parameters of the probability interval for the regression aiming missing data imputation.

A series of computing experiments performed with EM algorithm, forecast regression based imputation technique and some other approaches shows that it is possible to provide high quality results regarding correct processing of data with missing values. Usually it is difficult to predict which technique will provide better final result that is why it is recommended to use alternative computational procedures and select the best one for a specific case using known quality statistics. The combination of estimates computed with alternative techniques is also perspective approach to solving the problem of quality data imputation in the frames of DSS.

In future studies it is necessary to automate the process of data analysis directed towards its quality improvement with imputation of missing values. An automatic procedure supposes application of a variety of the techniques providing different intermediate results.

List of literature

1. *Svolba G.* Data Quality for Analytics Using SAS. – SAS Institute Inc., Cary, NC, 2012. – 340 p.
2. *Gogishvili P.* Determination of the vehicle location in case of incomplete GPS data // Інформаційні технології та комп'ютерна інженерія. – 2012. – № 3. – С. 19–23.
3. *Siddiqi N.* Credit Risk Scorecards: Developing and Implementing Intelligent Credit Scoring. – Hoboken, John Wiley & Sons, Inc., 2005. – 196 p.
4. *Owen M.* Tukey's Biweight Correlation and the Breakdown [Online]. – 2005. – Aavaliable: http://pages.pomona.edu/~jsh04747/Student%20Theses/MaryOwen10.pdf
5. *Breheny P.* Robust Regression [Online]. – Pomona: Pomona University, 2005. – 165 p. – Aavaliable: http://web.as.uky.edu/statistics/users/pbreheny/764-F11/notes/12-1.pdf
6. *Missing* value estimation for microarray data by bayesian principal component analysis and iterative local least squares [Online] / F. Shi, D. Zhang, J. Chen, H.R. Karimi // Math. Problems Eng. – 2013. – Article ID 162938. – Aavaliable: http://www.hindawi.com/journals/mpe/2013/162938/
7. *Marwala T.* Flexibly-bounded Rationality and Marginalization of Irrationality Theories for Decision Making [Online]. – 2014. – Avaliable: http://arxiv.org/ftp/arxiv/papers/1306/1306.2025.pdf
8. *McLachlan G.J.*, *Krishnan T.* The EM Algorithm and Extensions. – Hoboken: John Wiley & Sons Inc., 2008. – 359 p.

References

1. G. Svolba, *Data Quality for Analytics Using SAS*. Cary, NC: SAS Institute Inc., 2012, 340 p.
2. P. Gogishvili, “Determination of the vehicle location in case of incomplete GPS data”, *Informatsiyni Tekhnolohiyi ta Kompyuterna Inzheneriya,* no. 3, pp. 19–23, 2012 (in Russian).
3. N. Siddiqi, *Credit Risk Scorecards: Developing and Implementing Intelligent Credit Scoring*. Hoboken, John Wiley & Sons, Inc., 2005, 196 p.
4. M. Owen. (2005). *Tukey's Biweight Correlation and the Breakdown* [Online]. Avaliable: http://pages.pomona.edu/~jsh04747/Student%20Theses/MaryOwen10.pdf
5. P. Breheny, *Robust Regression* [Online]. Avaliable: http://web.as.uky.edu/statistics/users/pbreheny/764-F11/notes/12-1.pdf
6. F. Shi *et al.* (2013). *Missing Value Estimation for Microarray Data by Bayesian Principal Component Analysis and Iterative Local Least Squares*, [Online]. Avaliable: http://www.hindawi.com/journals/mpe/2013/162938/
7. T. Marwala. (2014). *Flexibly-bounded Rationality and Marginalization of Irrationality Theories for Decision Makin* [Online]. Avaliable: http://arxiv.org/ftp/arxiv/papers/1306/1306.2025.pdf
8. G.J. McLachlan and T. Krishnan. *The EM Algorithm and Extensions.* Hoboken: John Wiley & Sons, Inc., 2008, 359 p.

Н.В. Кузнєцова, П.І. Бідюк

ТЕХНОЛОГІЇ ІНТЕЛЕКТУАЛЬНИХ БІЗНЕС-ПЛАТФОРМ ДЛЯ ЗАПОВНЕННЯ ПРОПУСКІВ ДАНИХ

Проблематика. Належним чином спроектовані системи підтримки прийняття рішень для моделювання і прогнозування поведінки динамічних систем надають можливість врахування невизначеностей ймовірнісного, статистичного і структурного типів. Це сприяє підвищенню якості розроблюваних моделей та оцінок прогнозів.

Мета дослідження. Розглянути загальні причини втрати даних при розв’язанні задач їх статистичного аналізу; виконати категоризацію пропусків даних на кілька груп та виявити причини появи пропусків; запропонувати системну методологію аналізу невизначеностей та вибору методів заповнення пропусків; розглянути деякі популярні методи заповнення пропусків та можливості їх застосування.

Методика реалізації. Для розв’язання поставлених задач використано такі методи: підхід до категоризації пропусків даних з практичної та ділової точок зору з метою виявлення причин появи систематичних або випадкових втрат даних; сучасна методологія системного аналізу для встановлення причин появи невизначеностей та розв’язання задачі заповнення пропусків; процедури заповнення пропусків даних за допомогою дерев рішень; алгоритм ЕМ та підхід до заповнення пропусків за допомогою функцій прогнозування, що будуються на основі регресійних моделей.

Результати дослідження. Основними результатами дослідження є такі: категоризація пропущених даних на групи; розробка системної методології аналізу причини появи невизначеностей та розв’язання задачі заповнення пропусків; аналіз процедур заповнення пропусків за допомогою дерев рішень, алгоритму ЕМ та регресійних моделей. Наведено ілюстрацію застосування деяких перспективних методів заповнення пропусків.

Висновки. Запропоновано методику заповнення пропусків даних із шести кроків, яка підкреслює, що вибір коректного методу заповнення тісно пов’язаний із докладним аналізом причин появи пропусків. Результати заповнення пропусків іноді істотно відрізняються від фактичних даних, а тому їх необхідно згладжувати або навіть видаляти з вибірки внаслідок їх некоректності. У таких випадках необхідно використовувати ймовірнісно-регресійні процедури, які надають можливість визначати параметри ймовірнісних інтервалів регресії при генеруванні кандидатів на заповнення. Обчислювальні експерименти, виконані із застосуванням алгоритму ЕМ, оцінок прогнозів, отриманих на основі регресійних моделей та деяких інших методів, свідчать про те, що існують можливості для одержання високоякісних результатів обробки даних з пропусками.

Ключові слова: невизначеності, які трапляються в обробці даних; заповнення пропусків; системний підхід; системи підтримки прийняття рішень.

Н.В. Кузнецова, П.И. Бидюк

ТЕХНОЛОГИИ ИНТЕЛЛЕКТУАЛЬНЫХ БИЗНЕС-ПЛАТФОРМ ДЛЯ ЗАПОЛНЕНИЯ ПРОПУСКОВ ДАННЫХ

Проблематика. Системы поддержки принятия решений, спроектированные в соответствии с современными требованиями для решения задач моделирования и прогнозирования поведения динамических систем, дают возможность учета неопределенностей вероятностного, статистического и структурного типов. Это приводит к повышению качества разрабатываемых моделей и оценок прогнозов.

Цель исследования. Рассмотреть общие причины потери данных при решении задач их статистического анализа; выполнить категоризацию пропусков данных на несколько групп и определить причины появления пропусков; предложить системную методологию анализа неопределенностей и выбора методов заполнения пропусков; рассмотреть некоторые популярные методы заполнения пропусков, а также возможности их применения.

Методика реализации. Для решения поставленных задач использованы такие методы: подход к категоризации пропусков данных с практической и деловой точек зрения с целью выявления причин появления систематических или случайных потерь данных; современная методология системного анализа для установления причин появления неопределенностей и решения задачи заполнения пропусков; процедуры заполнения пропусков данных с помощью деревьев решений; алгоритм ЕМ и подход к заполнению пропусков с помощью функций прогнозирования, которые строятся на основе регрессионных моделей.

Результаты исследования. Основными результатами исследования являются такие: категоризация пропущенных данных на группы; разработка системной методологии анализа причин появления неопределенностей и решение задачи заполнения пропусков; анализ процедур заполнения пропусков с помощью деревьев решений, алгоритма ЕМ и регрессионных моделей. Приведена иллюстрация применения некоторых перспективных методов заполнения пропусков.

Выводы. Предложена методика заполнения пропусков данных с шести шагов, в которой подчеркивается, что выбор корректного метода заполнения тесно связан с углубленным анализом причин появления пропусков. Результаты заполнения пропусков иногда существенно отличаются от фактических данных, а потому их необходимо сглаживать или даже удалять с выборки из-за их некорректности. В таких случаях необходимо использовать вероятностно-регрессионные процедуры, которые дают возможность определять параметры вероятностных интервалов регрессии в процессе генерирования кандидатов на заполнение. Вычислительные эксперименты, выполненные с использованием алгоритма ЕМ, а также оценок прогнозов, полученных на основе регрессионных моделей и некоторых других методов, свидетельствуют о том, что существуют возможности для получения высококачественных результатов обработки данных с пропусками.

Ключевые слова: неопределенности, встречающиеся в обработке данных; заполнение пропусков; системный подход; системы поддержки принятия решений.

Рекомендована Радою
Навчально-наукового комплексу
“Інститут прикладного системного
аналізу” НТУУ “КПІ”

Надійшла до редакції
15 червня 2015 року