Some Imputation Algorithms for Restoration of Missing Data

Vladimir Ryazanov

Institution of Russian Academy of Sciences Dorodnicyn Computing Centre of RAS,
Vavilov st. 40, 119333 Moscow, Russia
http://www.ccas.ru

Abstract. The problem of reconstructing the feature values in samples of objects given in terms of numerical features is considered. The three approaches, not involving the use of probability models and a priori information, are considered. The first approach is based on the organization of the iterative procedure for successive elaboration of missing values of attributes. In this case, the analysis of local information for each object with missing data is fulfilled. The second approach is based on solving an optimization problem. We calculate such previously unknown feature values for which there is maximum correspondence of metric relations between objects in subspaces of known partial values and found full descriptions. The third approach is based on solving a series of recognition tasks for each missing value. Comparisons of these approaches on simulated and real problems are presented.

Keywords: missing data, imputation, feature, pattern recognition, feature values restoration.

1 Introduction

Many problems in data mining can be written in the standard form. Let be given a sample $\{z_i, \bar{x}_i\}, i = 1, 2, ..., m$, $\bar{x}_i = (x_{i1}, x_{i2}, ..., x_{in})$ is the feature description of some object, $z_i, x_{ij} \in R$. We assume scalar z_i is defined by vector $\bar{x}_i$. It is necessary to calculate $z = f(\bar{x})$ by some vector $\bar{x}$. Here $z, x_j \in R$.

Here we can distinguish three specific tasks:

1. $z \in \{1, 2, ..., l\}$,$z_i, i = 1, 2, ..., m$, are known (supervised classification or recognition task);
2. $z \in \{1, 2, ..., l\}$, but the values $z_i, i = 1, 2, ..., m$, (and may be l) are unknown (unsupervised classification or clustering task);
3. $z \in (a, b)$, $z_i, i = 1, 2, ..., m$ are known (task of regression reconstruction).

In this paper we consider the case of missing data for some features (unknown feature values are denoted by Δ).

There are various approaches: taking into account the type of tasks (clustering, classification or regression), cases of training or classification , taking

C. San Martin and S.-W. Kim (Eds.): CIARP 2011, LNCS 7042, pp. 372–379, 2011.

into account additional a priori knowledge and hypotheses, the direct solution of problems with missing data or their decision after a preliminary gaps reconstruction . We consider the case when the problems are solved in the following two steps. First signs of recovering missing values in object descriptions. In the second phase is addressing these problems for a complete description, which already uses the standard well-known algorithms.

There are different approaches to solve the problem of reconstruction of missing feature values that are commonly referred to as marginalisation, imputation, and projection. In the case of marginalization or skipping incomplete objects, the incomplete objects in the dataset are discarded simply in order to create a new complete dataset [1]. In this case, you may lose a large amount of information. In the case of *Imputation approaches,* a value from the entire dataset to fill the missing attribute is estimated. The well-known imputation techniques are the mean of known values of the same feature in other instances, median, random [1, 2], the nearest neighbour method [3]. In [4], a partial imputation technique has been proposed. It consists of the imputation of missing data using complete objects in a small neighborhood of the incomplete ones. In [5], a new approach is proposed , using the entropy to estimate the missing values.

The *Projection methods* (or imputation by regression) realize the next idea. The feature space is reduced to one dimension less for each missing attribute. So, it is necessary to compute a special classifier or regression function in the reduced space. Usually, complete objects of the training set are used to build the optimal classifier\regression. In [6], the imputation technique using support vector regressions (SVR) is studied and compared with some well-known ones. The results showed the high precision obtained by SVR technique with regards to the mean, median, of the nearest neighbor techniques.

It is well known and reliable algorithm for filling gaps by maximum likelihood (EM algorithm) [1, 2]. A disadvantage is the low rate of convergence, if missed a lot of data. Probably, there are a lot of local optimal solutions. It is assumed a reasonable probabilistic model of classification or regression.

In this article we propose three algorithms for the restoration feature values according to the training samples, based on attempts to implement the following principles:

- use all the objects of training sample, regardless of the number of existing gaps;
- do not use any probabilistic assumptions about the data set;
- background information is only sample data;
- features in general are not independent.

Initial information is training sample of objects $X = \{\bar{x}_1, \bar{x}_2, ..., \bar{x}_m\}$. We assume that $x_{ij} \in \{M_j, \Delta\}, M_j \subseteq R$. Unknown feature values x_{ij} are denoted as Δ. Let the set of pairs of indexes J specifies all unknown values of attributes of the objects of training sample $J = \{\langle i, j\rangle : x_{ij} = \Delta\}$. Region $M_j, j = 1, 2, ..., n$, of permissible values of each feature is a finite set, which is determined by a given sample.

The task of reconstruction of unknown feature values is to find a sample $X^* = \{\bar{x}_1^*, \bar{x}_2^*, ..., \bar{x}_m^*\}$ of complete descriptions $\bar{x}_i^* = (x_{i1}^*, x_{i2}^*, ..., x_{in}^*)$, $x_{ij}^* = \begin{cases} x_{ij}, & x_{ij} \neq \Delta, \\ \in M_j, & x_{ij} = \Delta, \end{cases}$, "the most corresponding" sample given a partial descriptions $\bar{X}$. This "best match" can be defined explicitly or not explicitly. We consider the following three approaches and specific algorithms.

2 Local Method for Reconstructing Feature Values

First, all the unknown feature values are filled with random numbers from the range of admissible values of variable $x_{ij} \in M_j, j = 1, 2, ..., n$. Next, the unknown values sequentially modified by a combination of method k-nearest neighbor and shift procedure. Let be given the metric in the space of the feature descriptions.

Step 0. Initializing random $x_{ij}^{(0)} \in M_j, \forall \langle i, j \rangle \in J$. Obtain the full descriptions. If $\langle i, j \rangle \in J$, let $x_{ij}^{(0)*}$ is an average value of feature j over the k nearest neighbors of $\bar{x}_i$. Then define $x_{ij}^1 = x_{ij}^{(0)} + \theta(x_{ij}^{(0)*} - x_{ij}^{(0)}), \forall \langle i, j \rangle \in J$, $x_{ij}^{(1)} = x_{ij}^{(0)}, \forall \langle i, j \rangle \notin J$. Here $0 < \theta \leq 1$.

Step t=1,2,.... We have $\bar{x}_i^{(t-1)} = (x_{i1}^{(t-1)}, ..., x_{in}^{(t-1)})$. For each pair $\langle i, j \rangle \in J$, the $x_{ij}^{(t-1)*}$ is calculated as the average value of feature j over the k- nearest neighbors of the object $x_i^{(t-1)}$. Then define $x_{ij}^{(t)} = x_{ij}^{(t-1)} + \theta(x_{ij}^{(t-1)*} - x_{ij}^{(t-1)})$, $\forall \langle i, j \rangle \in J$, $x_{ij}^{(t)} = x_{ij}^{(t-1)}, \forall \langle i, j \rangle \notin J$. Step is repeated, if not satisfied the stopping criterion. Otherwise, the restoration of gaps is finished.

Stopping criterion: the maximum number of iterations N, $\sum_{\langle i,j \rangle \in J} \left| x_{ij}^{(t)} - x_{ij}^{(t-1)} \right|^2 \leq \varepsilon$, etc. Finally, we put $x_{ij}^{(final)}, \forall \langle i, j \rangle \in J$ the nearest value from M_j.

3 Optimization Method for Reconstructing Feature Values

The essence of this approach is that missing values should take such values for which the metric relationships between objects in space of "full descriptions" as would correspond to metric relations in the subspaces of known "partial descriptions".

Let $\bar{x}_i, \bar{x}_j$ is a pair of training objects. We introduce the notation: $\Omega_i^0 = \{t : x_{it} \neq \Delta\}$, $\Omega_i^1 = \{t : x_{it} = \Delta\}$. Let $\Omega_{ij}^{00} = \Omega_i^0 \bigcap \Omega_j^0$, $\Omega_{ij}^{01} = \Omega_i^0 \bigcap \Omega_j^1$, $\Omega_{ij}^{10} = \Omega_i^1 \bigcap \Omega_j^0$, $\Omega_{ij}^{11} = \Omega_i^1 \bigcap \Omega_j^1$. We will use the Euclidean metric $\rho^2(\bar{x}_i, \bar{x}_j) = \left(\sum_{t \in \Omega_{ij}^{00}} (x_{it} - x_{jt})^2 + \sum_{t \in \Omega_{ij}^{01}} (x_{it} - y_{jt})^2 + \sum_{t \in \Omega_{ij}^{10}} (y_{it} - x_{jt})^2 + \sum_{t \in \Omega_{ij}^{11}} (y_{it} - y_{jt})^2 \right)$. Here and below, for convenience, the unknown values of features x_{it} are replaced by the parameters y_{it}: $\{\langle i, j \rangle \in J\}$ for all pairs of indexes.

We will consider the next distances in the feature subspaces.

$\rho^{+}(\bar{x}_i, \bar{x}_j) = \left(\sum_{t \in \Omega_{ij}^{00}} (x_{it} - x_{jt})^2\right)^{\frac{1}{2}}$ is a distance in the subspace in which the values of the features of both objects are known.

$\rho^{++}(\bar{x}_i, \bar{x}_j) = \left(\sum_{t \in \Omega_{ij}^{00}} (x_{it} - x_{jt})^2 + \sum_{t \in \Omega_{ij}^{01}} (x_{it} - y_{jt})^2 + \sum_{t \in \Omega_{ij}^{10}} (y_{it} - x_{jt})^2\right)^{\frac{1}{2}}$ is a distance in the subspace, in which the feature values are known at least for one object.

We consider the following two criteria of filling gaps quality as a function of the unknown values of features:

$$\Phi(\langle y_{ij} \rangle) = \sum_{\substack{i,j=1 \\ i>j}}^{m} (\rho(\bar{x}_i, \bar{x}_j) - N_{ij}^{+} \rho^{+}(\bar{x}_i, \bar{x}_j))^2,$$

$$F(\langle y_{ij} \rangle) = \sum_{\substack{i,j=1 \\ i>j}}^{m} (\rho(\bar{x}_i, \bar{x}_j) - N_{ij}^{++} \rho^{++}(\bar{x}_i, \bar{x}_j))^2.$$

Here N_{ij}^{+}, $N_{ij}^{++}$are chosen according to one of the following ways:

1.a $N_{ij}^{+} = 1$, 1.b. $N_{ij}^{+} = \frac{n}{|\Omega_{ij}^{00}|}$ (if $|\Omega_{ij}^{00}| = 0$ put $\rho(\bar{x}_i, \bar{x}_j) - N_{ij}^{+} \rho^{+}(\bar{x}_i, \bar{x}_j) = 0$).
2.a. $N_{ij}^{++} = 1$, 2.b. $N_{ij}^{++} = \frac{n}{|\Omega_{ij}^{00} + \Omega_{ij}^{01} + \Omega_{ij}^{10}|}$

(if $|\Omega_{ij}^{00} + \Omega_{ij}^{01} + \Omega_{ij}^{10}| = 0$ put $\rho(\bar{x}_i, \bar{x}_j) - N_{ij}^{++} \rho^{++}(\bar{x}_i, \bar{x}_j) = 0$).

Gradient of the first criterion is as follows

$$\frac{\partial \Phi(\langle y_{ij} \rangle)}{\partial y_{\alpha\beta}} = 2 \sum_{\substack{i \neq \alpha, \\ \rho^{+}(\bar{x}_i, \bar{x}_\alpha) > 0}} \frac{(\rho(\bar{x}_i, \bar{x}_\alpha) - N_{i\alpha}^{+} \rho^{+}(\bar{x}_i, \bar{x}_\alpha))}{\rho(\bar{x}_i, \bar{x}_\alpha)} (y_{\alpha\beta} - x_{i\beta}),$$

Gradient for $F(\langle y_{ij} \rangle)$ is the analogy one.

Then we apply the method of steepest descent with constraints
$y_{ij} \in [\min_{t=1,2,\ldots,m} x_{tj}, \max_{t=1,2,\ldots,m} x_{tj}]$.

In the local and optimization approaches, we do not distinguish between numeric and discrete features. After restoring the values of features, we put $x_{ij}^{*} = a_{sj} : a_{sj} \in M_j, \left|a_{sj} - x_{ij}^{(final)}\right| = \min_{\substack{t=1,2,\ldots,m \\ <t,j> \notin J}} \left|a_{tj} - x_{ij}^{(final)}\right|$. In the case of two possible solutions, we take one from them which has a higher frequency on the training data.

4 Restoration of Feature Values as the Solution of Recognition Problem

Meaningful task is to assign these numerical values for objects with a gaps, which are the most "agreed" with known features of the object. The reconstruction

problem is solved sequentially for each pair $\langle i, j\rangle \in J$ as a special recognition task. Let for a object $\bar{x}_i$ from the sample X value x_{ij} is unknown. For simplicity, we further denote $\bar{y} = \bar{x}_i, \bar{y} = (y_1, y_2, ..., y_n)$, $\Omega_i = \{j_1, j_2, ...j_\tau\}, \Theta_i = \{k_1, k_2, ..., k_\sigma\} = \{1, 2, ..., n\}\backslash\Omega_i$, $x_{ij} = \Delta$, $\forall j \in \Omega_i$. Denote $M_j = \{a_1, a_2, ..., a_N\}$ as the set of all possible values of $j-$th feature. It is calculated by known data. Let $a = a_1 < a_2 < ... < a_N = b$.

The general algorithm consists in solving of $[\log_2 N] + 1$ dichotomous recognition tasks. It was used the estimation calculation algorithm, based on voting over support sets of a given power [7]. This algorithm reflects the correlation properties between features and doesn't use any training.

1. There is a set of numbers $a = a_1 < a_2 < ... < a_N = b$. We consider two classes: $K_1 = \{\bar{x}|a \le x_j \le a_{\left[\frac{N}{2}\right]}\}$, $\tilde{K_1} = K_1 \bigcap X$, $K_2 = \{\bar{x}|a_{\left[\frac{N}{2}\right]} < x_j \le b\}$, $\tilde{K_2} = K_2 \bigcap X$.
2. Estimate $\Gamma_t(\bar{y}) = \sum_{\bar{x}_\lambda \in \tilde{K}_t} C^k_{d(\bar{x}_\lambda, \bar{y})}$ for class $K_t, t = 1, 2$ (degree of membership of an object $\bar{y}$ to class $K_t, t = 1, 2$) is computed. Here $d(\bar{x}_\lambda, \bar{y}) = |\{\beta : |y_\beta - x_{\lambda\beta}| \le \varepsilon_\beta\}, \beta \in \Theta_i \bigcap \Theta_\lambda|$, $1 \le k \le n$ is an integer (control input parameter), $\varepsilon_\beta = \frac{2}{|h_\beta|(|h_\beta|-1)} \sum_{\bar{x}_u, \bar{x}_v \in \tilde{K}_1 \bigcup \tilde{K}_2, u>v, x_{u\beta}, x_{v\beta} \ne \Delta} |x_{u\beta} - x_{v\beta}|$, $h_\beta = \left|\{\bar{x}_a \in \tilde{K_1} \bigcup \tilde{K_2} : x_{a\beta} \ne \Delta\}\right|$.
3. If $\Gamma_1(\bar{y}) \ge \Gamma_2(\bar{y})$ put $\bar{y} \in K_1$, otherwise $\bar{y} \in K_2$.
4. If, the class to which $\bar{y}$ is assigned contains only one element $\bar{x}_a$, then we put $y_j = x_{aj}$. The task of restoring the value x_{ij} is considered to be resolved. Otherwise, the transition at point 1 and process is repeated with respect to the set $a_1 < a_2 < ... < a_{\left[\frac{N}{2}\right]}$ (if $\bar{y}$ assigned in class K_1) or relative to the set $a_{\left[\frac{N}{2}\right]+1} < a_{\left[\frac{N}{2}\right]+2} < ... < a_N$ (if $\bar{y}$ assigned in class K_2). It is clear that no more than $[\log_2 N] + 1$ steps, we obtain the first situation 4.

To calculate estimates for the classes one can use other ways of calculating estimates [7].

5 The Results of Numerical Experiments

This section presents the initial results of the application and comparison of some different feature values restoration techniques. Two models of the creation of data gaps have been considered.

In the first model in each training object $\alpha\%$ of feature values were considered missing. This selection was performed randomly in a uniform distribution law. In the second model, $\alpha\%$ of elements of training set were considered as missing. The choice of these pairs "object-feature" also was performed randomly according uniform distribution law. Experiments were carried out as follows. According to the original training set, the training samples with gaps were formed by the first or second model. The samples of incomplete feature descriptions were restored by the algorithm mean substitution, and the algorithms proposed in this paper. After that, for all tables were solved the supervised classification (recognition)

problems using different algorithms. We used the implementation of algorithms in a software system " Recognition" [8]. Experiments were conducted with model data and with one practical problem.

As a model task, a mixture of two normal distributions has been considered. Training and the control data consisted of 200 vectors, 100 ones from each class. The vectors consisted of values of 10 independent features. Features of the first (second) class are normally distributed according normal distributions with parameters $a = 0, \sigma^2 = 9$ (respectively $a = 5, \sigma^2 = 9$), where the a is an expectation, σ^2 is a variance. Transformations of training and control data in the samples with gaps were run with α= 35. Restoration of training and control samples were carried out independently. Fig. 1 shows the visualization of control sample. Black and gray dots correspond to the first and second classes, respectively. They are displaying objects from a R^{10}on a plane with maximum preservation of metric relations between objects in R^{10} [9].

Fig. 1. Visualization of the control sample in a simple model task

Tables 1 and 2 present the recognition results to a control sample and its modifications by various algorithms.

Recognition accuracy was estimated as the percentage of correctly recognized objects of control sample. When training of the various algorithms used standard values of their control parameters. The task of choosing the optimal parameters of the algorithms for training was not considered. So, the results of different algorithms for solving same tasks are very different. The rows of tables presents the results of recognition of different algorithms: LM – "linear machine" [9], k –neighbors – "k-nearest neighbors" [9], AEC – "estimation calculation algorithms" [7], LDF – "Fisher linear discriminator"[9], LR - "voting algorithm over sets of logical regularities of classes [10].

Each column of the table presents the results of recognition of different modifications of the original checklist: the source table, the method of mean substitution, applications of methods 1, 2, 3, that denote the local, optimization and based on pattern recognition tasks solving methods. In the local method we used the values of parameters $\theta = 0.8$, $N = 50$, $k = 5$, in an optimization algorithm has been used functional $\Phi(\langle y_{ij} \rangle)$ for $N_{ij}^{+} = 1$.

Table 1. Recognition of simulated data for the first model of gaps creation

recogn. method \ table	source table	mean substitution	method 1	method 2	method 3
LM	84.5	85.0	84.5	86.0	84.5
k – neighbors	87.0	82.0	82.5	83.0	81.5
AEC	86.0	73.0	79.5	75.5	80.0
LDF	86.0	81.0	82.5	81.5	77.5
LR	85.0	72.5	81.0	78.5	78.0

Table 2. Recognition of simulated data for the second model of gaps creation

recogn. method \ table	source table	mean substitution	method 1	method 2	method 3
LM	84.5	85.5	81.5	86.0	84.5
k – neighbors	87.0	82.5	85.5	81.5	85.0
AEC	86.0	75.0	78.0	78.5	83.0
LDF	86.0	79.0	81.5	79.5	82.0
LR	85.0	69.9	77.5	79.5	81.5

6 Conclusion

There was considered a model task that has been created on the basis of three normal distributions having linearly inseparable centers of the classes. Besides, here were used some other recognition algorithms (neural network with back propagation training [11], binary decision trees [9], SVM [12], multiplicative neural network [13]). As a test task,we examined a sample of patients with complaints of chest pain from Heart Disease Databases (Hungarian Institute of Cardiology. Budapest: Andras Janosi, M.D., University Hospital, Zurich, Switzerland: William Steinbrunn, M.D., University Hospital, Basel, Switzerland: Matthias Pfisterer, M.D., V.A. Medical Center, Long Beach and Cleveland Clinic Foundation: Robert Detrano, M.D., Ph.D.). The results were similar to those of considered earlier.

For a more accurate evaluation of the proposed approaches and their comparison necessary to carry out a large series of experiments both on real and simulated data, and various models of missing data modelling. Nevertheless, these preliminary calculations confirm some a priori expectations. Local averaging of characteristics (method 1) is better than the average for the full sample. Method 2 showed good results, but apparently it will be inefficient for problems with a large number of gaps. In method 3 is used an algorithm AEC. The calculation of the degree of affiliation $\Gamma_t(\bar{y})$ for object $\bar{y}$ to a certain class K_t is based on a comparison $\bar{y}$ with each $\bar{x}_\lambda \in K_t$. Comparison takes place on different subsets of

features in the maximum feature subspace where $\bar{y}$ and $\bar{x}_\lambda$ have no gaps. This expresses the fact of existance of dependencies between features (see [7]).

The total ratio of the first places of compared methods is 4:7:16:9. Methods 1-3 show generally higher results. In any case, the creation of new algorithms for reconstruction of unknown feature values is important. Having a set of different recovery algorithms, we improve the chances of a more exact solution of classification problems.

Acknowledgments. This work was supported by RAS Presidium programs number 14 and "Basic Sciences - Medicine, Program number 2 of Department of Mathematical Sciences of RAS, RFBR 09-01-00409, 10-01-90015 Bel_a, 10-01-90419 Ukr_a.

References

1. Little, R.J.A., Rubin, D.B.: Statistical Analysis with Missing Data. Wiley, New York (1987)
2. Zloba, E.: Statistical methods of reproducing of missing data. J. Computer Modelling & New Technologies 6(1), 51–61 (2002)
3. Morin, R.L., Raeside, D.E.: A reappraisal of distance-weighted k-nearest neighbor classification for pattern recognition with missing data. IEEE Transactions on Systems, Man and Cybernetics, 241–243 (1981)
4. Zhang, S.: Parimputation: From imputation and null-imputation to partially imputation. IEEE Intelligent Informatics Bulletin 9(1), 32–38 (2008)
5. Delavallade, T., Dang, T.H.: Using Entropy to Impute Missing Data in a Classification Task. In: IEEE International Conference on Fuzzy Systems, London, pp. 1–6 (2007)
6. Honghai, F., Guoshun, C., Cheng, Y., Bingru, Y., Yumei, C.: A SVM Regression Based Approach to Filling in Missing Values. In: Khosla, R., Howlett, R.J., Jain, L.C. (eds.) KES 2005. LNCS (LNAI), vol. 3683, pp. 581–587. Springer, Heidelberg (2005)
7. Zhuravlev, Y.I., Nikiforov, V.V.: Recognition Algorithms based on Estimate Evaluation. J. Kibernetika 3, 1–11 (1971) (in Russian)
8. Zhuravlev, Y.I., Ryazanov, V.V., Senko, O.V.: Recognition. Mathematical methods. Programm. System. Applications, Fazis, Moscow (2006) (in Russian)
9. Duda, R.O., Hart, P.E., Stork, D.G.: Pattern Classification, 2nd edn. Wiley interscience (2001)
10. Ryazanov, V.V.: Logical Regularities in Pattern Recognition (Parametric Approach). Computational Mathematics and Mathematical Physics 47(10), 1720–1735 (2007); ©Pleiades Publishing, Ltd., Original Russian Text ©V.V. Ryazanov, published in Zhurnal Vychislitel'noi Matematiki i Matematicheskoi Fiziki 47(10), 1793–1808 (2007)
11. Fausett, L.: Fundamentals of Neural Networks. Prentice-Hall (1994)
12. Vapnik, V.: The Nature of Statistical Learning Theory. Springer, Heidelberg (1995)
13. Ishodzhanov, T.R., Ryazanov, V.V.: A gradient search for logical regularities of classes with a linear dependence. In: 14th All-Russian Conference on Mathematical Methods for Pattern Recognition: 14 All-Russian Conference, pp. 123–124. MAKS Press, Vladimir region (2009)