CLASSIFICATION OF INCOMPLETE DATA

Vasily V.Ryazanov, Vladimir V.Ryazanov

Abstract: *The problem of reconstructing the feature values in samples of objects given in terms of numerical features is considered. A numerical study of different approaches of solving this problem on one model and three practical problems at different levels of data incompleteness is held. A modification of the model calculation of estimates, not requiring metrics for signs, is suggested. The advantage of a local and recognition approaches over filling gaps with sample averages is shown.*

Keywords: *data mining, missing data, classification, pattern recognition.*

ACM Classification Keywords: *I.5 Computing Methodologies – Pattern recognition*

Introduction

In many problems of data analysis, supervised or unsupervised classification, regression problems as the study data samples $X = \{\mathbf{x}_i, i = 1,2,...,m\}$ are used, where $\mathbf{x}_i = (x_{i1}, x_{i2}, ..., x_{in})$ - is the feature description of some object, $x_{ij} \in R$.

In this paper, we consider the case when part of feature values x_{ij} of objects is unknown (unknown values are denoted Δ). Various approaches to restore the feature values are possible. Some of them take into account the kind of task (clustering, classification or regression), probabilistic characteristics of features, additional prior knowledge and hypotheses.

Currently, there are different approaches to solve the problem of restoration missing data which are conventionally divided into «marginalization», «imputation» и «projection».

In case of «marginalization» or «skipping incomplete objects» incomplete descriptions are simply excluded from the sample. As a result a new sample of complete descriptions is formed [Little, Rubin, 1987]. In this case, obviously, a lot of information can be lost.

In case of «imputation» the most suitable value for a gap is evaluated from the entire sample. General methodology of «imputation» uses the sample means, median, random choice [Little, Rubin, 1987; Zloba, 2002]. In paper [Morin, Raeside, 1981] nearest neighbor algorithm was used. Recently Zhang [4] proposed a method of «partial imputation». Unknown data is estimated by full description of the object in a small neighborhood of incomplete descriptions. Delavallade and Ha [Delavallade, Dang 2007] proposed a new approach which uses entropy to estimate unknown values.

In «projection»-methods (or «imputation by regression») feature space is reduced by one for each unknown characteristic values. This requires a special construction of the classifier in the reduced space. Usually full description from the training set for construction of an optimal hyperplane are used. In [Honghaj etc., 2005], Honghai etc. investigated «imputation» - approach with the use of support vector machines (SVM) and conducted its comparison with the replacement of the unknown values with mean, median, average of the next two values, and the values of the nearest neighbor method. The results show that the SVM method showed the highest accuracy in comparison to other methods. The algorithm of filling the gaps with maximum likelihood is well-known and reliable (EM algorithm) [Little, Rubin, 1987]. The disadvantage of the method is low rate of convergence, if a

lot of data is missed. In solving the problem locally optimal solutions are found. A probability model of forming data is assumed reasonable.

In paper [Ryazanov, 2011] three approaches to the restoration of feature values are considered. The first is based on the organization of the iterative procedure of the sequential refinement of missing feature values. Herewith, the analysis of local information for each object with gaps is performed. The second approach is based on solving the optimization problem. Previously unknown feature values are found according to the rule that there is maximal conformity of metric relations between objects in subspaces of known values and their found spaces of full descriptions. In the third approach, each missing feature value is found by solving a series of problems of recognition.

In this paper, a numerical study of the first and third approaches in various model and real practical problems is conducted. Same time a new model to detect unknown feature values is suggested. This model doesn't require presence of metrics for features. In the numerical experiments for comparison the «means» method is used (replacement of unknown feature values with sample averages).

During restoring unknown feature values the following ideas are implemented:

- all training data objects regardless of the number of gaps in them are used;
- no probabilistic assumptions about the data sample are used;
- the only initial information is sample data;
- features in general are not independent and metric.

We assume that $x_{ij} \in \{M_j, \Delta\}, M_j \subseteq R$, R - set of real numbers with the metric $\rho(a,b) = |a-b|$. M_j - finite set of possible values of $j-$ feature, which consists of all of its values recorded at the training data. Let the set of pairs of indices J sets all unknown feature values of the objects of the training set $J = \{\langle i, j\rangle, i = 1,2,...,m, j = 1,2,...,n : x_{ij} = \Delta\}$. The task of restoration of unknown feature values is to find a sample $X^* = \{\mathbf{x}^*_1, \mathbf{x}^*_2, ..., \mathbf{x}^*_m\}$ of full descriptions $\mathbf{x}^*_i = (x^*_{i1}, x^*_{i2}, ..., x^*_{in})$, $x^*_{ij} = \begin{cases} x_{ij}, & x_{ij} \neq \Delta, \\ \in M_j, & x_{ij} = \Delta, \end{cases}$ which «maximum corresponds» to sample set of partial descriptions X.

1. Local method of restoration feature values

The idea is simple. Unknown values of feature should be similar to the known values of a small neighborhood of the object. Firstly, all unknown feature values are filled with random numbers from the feature tolerance range $x_{ij} \in M_j, j = 1,2,...,n$, or are filled based on some a priori considerations. Then unknown values are sequentially modified by a combination of k - neighbors method and shift procedure.

Fix a metric in the space of feature descriptions.

Step 0. Initialize the initial $x_{ij}^{(0)} \in M_j$, $\forall\langle i, j\rangle \in J$. Obtain a table full descriptions.

Step t=1,2,.... We have $\mathbf{x}_i^{(t-1)} = (x_{i1}^{(t-1)}, ..., x_{in}^{(t-1)})$. For each pair $\langle i, j\rangle \in J$ calculate $x_{ij}^{(t-1)}*$, as the average value of the feature with index j of k neighbors of $\mathbf{x}_i^{(t-1)}$ object. Set $x_{ij}^{(t)} = x_{ij}^{(t-1)} + \theta(x_{ij}^{(t-1)}* - x_{ij}^{(t-1)}), \forall\langle i, j\rangle \in J$, $x_{ij}^{(t)} = x_{ij}^{(t-1)}, \forall\langle i, j\rangle \notin J$. Here k is integer, $1 \le k \le m-1, 0 < \theta \le 1$ - algorithm parameters. The step is repeated, until the stopping criterion is executed.

As a stopping criterion are used the followings: the maximum number of iterations N , condition $\sum_{\langle i,j\rangle\in J}\left|x_{ij}^{(t)}-x_{ij}^{(t-1)}\right|^2\le\varepsilon$, etc.

Choice of metric, parameters θ , N , ε , k , initial values of the unknown features is made by the user.

2. Optimization method of restoration feature values

The essence of this approach is that missing values should take such values, with which the metric relationships between objects in space of «full descriptions» maximally correspond to the metric relationship in subspaces of «known descriptions» or «more known descriptions».

Let $\mathbf{x}_i,\mathbf{x}_j$ - any pair of training table rows. Introduce notation:

$\Omega_i^0=\{t:x_{it}\neq\Delta\}$, $\Omega_i^1=\{t:x_{it}=\Delta\}$. Set $\Omega_{ij}^{00}=\Omega_i^0\cap\Omega_j^0$, $\Omega_{ij}^{01}=\Omega_i^0\cap\Omega_j^1$, $\Omega_{ij}^{10}=\Omega_i^1\cap\Omega_j^0$, $\Omega_{ij}^{11}=\Omega_i^1\cap\Omega_j^1$.

The Euclidean metric is considered:

$$\rho(\mathbf{x}_i,\mathbf{x}_j)=\left(\sum_{t\in\Omega_{ij}^{00}}(x_{it}-x_{jt})^2+\sum_{t\in\Omega_{ij}^{01}}(x_{it}-y_{jt})^2+\sum_{t\in\Omega_{ij}^{10}}(y_{it}-x_{jt})^2+\sum_{t\in\Omega_{ij}^{11}}(y_{it}-y_{jt})^2\right)^{\frac{1}{2}},$$ where, for convenience, the unknown values of x_{it} features are replaced with the variables y_{it} .

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

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

In [Ryazanov, 2011] the problem of minimizing the following two criteria of filling the gaps was considered:

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

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

This approach is visual and natural, numerical experiments have confirmed its prospects. However, there are significant difficulties in its implementation, and adaptation to the data parameters. These optimization tasks have multiple extremums, so the choice of the initial approximation and analysis of local optima is the subject of a separate study. In addition, with a large number of gaps optimization task has a greater dimension and it should be taken into account in the optimization method. It seems that a practical method for optimizing should significantly depend on the degree of data incompleteness.

3. Restoring the features as a solution the problem of recognition

The task of restoring signs is solved sequentially for each pair $\langle i,j\rangle\in J$ as a special recognition task. Let for object $\mathbf{x}_i$ from sample X value x_{ij} is unknown. Then we denote,

$\Omega_i = \{j_1, j_2, ...j_\tau\}, \Theta_i = \{k_1, k_2, ..., k_\sigma\} = \{1, 2, ..., n\} \setminus \Omega_i$, where Ω_i - set of numbers from $\{1,2,...,n\}$, for which $x_{ij} = \Delta$. One considers that $M_j = \{a_1, a_2, ..., a_N\}$, $a = a_1 < a_2 < ... < a_N = b$. The main task is to recognize a_t, which will be the true value for x_{ij}.

General algorithm consists of solving $\lceil \log_2 N \rceil + 1$ dichotomous recognition problems. Assume we have some standard recognition algorithm A, which according to training set X of features descriptions solves the problem of classification with two classes K_1, K_2: $A : \mathbf{x} \to \mathrm{K}(\mathbf{x})$, where $\mathrm{K}(\mathbf{x}) \in \{K_1, K_2\}$, $\mathbf{x}$ - any object to be recognized. Let's describe the general scheme of the reconstruction algorithm of feature values, based on the solution of dichotomous recognition problems.

3.1. General algorithm

1. We have a set of numbers $a = a_1 < a_2 < ... < a_N = b$. Consider two classes:

$K_1 = \{\mathbf{x} \mid a \le x_j \le a_{\left\lfloor \frac{N}{2} \right\rfloor}\}$,

$\widetilde{K}_1 = K_1 \cap X$, $K_2 = \{\mathbf{x} \mid a_{\left\lfloor \frac{N}{2} \right\rfloor} < x_j \le b\}$, $\widetilde{K}_2 = K_2 \cap X$.

2. Based on the training set $\{\widetilde{K}_1, \widetilde{K}_2\}$ recognition algorithm $A_1 : \mathbf{x} \to \mathrm{K}(\mathbf{x})$, $\mathrm{K}(\mathbf{x}) \in \{K_1, K_2\}$ is constructed.

3. The task of recognition $\mathbf{x}$ is solved with classes K_1, K_2.

4. If the training set of class $\mathrm{K}(\mathbf{x})$ consists only of objects $\mathbf{x}_\alpha$, for which $x_{\alpha t} = a_t$, we set $x_{ij} = a_t$ and the feature value is considered to be calculated. Otherwise $\mathrm{K}(\mathbf{x})$ (and corresponding training objects) analogously 1) is divided into two sets of values $a_1, a_2, ..., a_N$, go to step 2) and the process is repeated for new classes. It is clear that no more than after $\lceil \log_2 N \rceil + 1$ steps, we obtain the solution.

Let each feature has $N_j, j = 1,2,...,n$ different values. Then one needs to solve no more than $\sum_{j=1}^{n} (\lceil \log_2 N_j \rceil + 1)$ training tasks. In practical calculations was used a model of calculating estimates, which does not require prior solutions of laborious training tasks.

3.2. Algorithms for calculating estimates with non-metric proximity function

Considered further modification of the model calculation of estimates [Zhuravlev, Nikiforov, 1971] can be used for object recognition with ordinal features by changing the proximity function. The set R is the set of real numbers. Consider the proximity function of recognized $\mathbf{x}$ comparatively to the pair of objects $\mathbf{x}_t$ и $\mathbf{x}_\tau$:

$$B_\Omega(\mathbf{x}_t, \mathbf{x}, \mathbf{x}_\tau) = \begin{cases} 1, & (x_{t\beta} \le x_\beta \le x_{\tau\beta}) \vee (x_{\tau\beta} \le x_\beta \le x_{t\beta}), \forall \beta \in \Omega, \\ 0, & \text{otherwise.} \end{cases}$$

To calculate the estimates the expression (1) is considered:

$$\tilde{\Gamma}_j(\mathbf{x}) = \frac{2}{\left|\tilde{K}_j\right|\left(\left|\tilde{K}_j\right| - 1\right)} \sum_{\mathbf{x}_t, \mathbf{x}_\tau \in \tilde{K}_j, t > \tau} \sum_{\Omega \in \Omega_A} B_\Omega(\mathbf{x}_t, \mathbf{x}, \mathbf{x}_\tau) \qquad (1)$$

It can be shown (similar to standard proximity functions [Zhuravlev, Nikiforov, 1971]), that there is a formula of effective computing estimates (1):

$$\tilde{\Gamma}_j(\mathbf{x}) = \frac{2}{\left|\tilde{K}_j\right|\left(\left|\tilde{K}_j\right|-1\right)} \sum_{\mathbf{x}_t,\mathbf{x}_\tau \in \tilde{K}_j, t>\tau} C^k_{\tilde{d}(\mathbf{x}_t,\mathbf{x},\mathbf{x}_\tau)}, \quad (2)$$

where $d(\mathbf{x}_t, \mathbf{x}, \mathbf{x}_\tau) = \left|\{\beta:(x_{t\beta} \le x_\beta \le x_{\tau\beta}) \vee (x_{\tau\beta} \le x_\beta \le x_{t\beta})\}, \beta = 1,2,...,n\right|$.

Natural limitation of the model is $\left|\tilde{K}_j\right| \ge 2$.

Note that the model of computing estimates in task of recognition unknown feature value x_{ij} can be used «directly», without assigning any values to other unknown feature values in the learning table. For this, during calculation of the quantities $d(\mathbf{x}_\alpha, \mathbf{x}_t)$ (thereafter, $d(\mathbf{x}_t, \mathbf{x}_\alpha, \mathbf{x}_\tau)$) it is sufficient to consider the cases $\beta \in \Theta_\alpha \cap \Theta_t$ (thereafter, $\beta \in \Theta_\alpha \cap \Theta_t \cap \Theta_\tau$). During solving the problems of recognition training data the calculation of any parameters is not required.

4. Numerical results

This section presents the results of numerical experiments on simulated data and two real practical problems. Because in fact all methods of reconstructing feature values require initial values, as initial values sample means for known feature values were used («mean» or «average» method [Little, Rubin, 1987]). This choice was driven by good results of given initial approximation and ability to compare the recognition results with the original averages. In addition, with a large number of gaps, random selection of these gaps results in a large scatter of reconstructed values and, therefore, leads to new task of choosing found values.

The experiments for each task were performed in the following way. Based on the training sample, sample with gaps was built. Each line of a new sample has $\alpha\%$ unknown features. This choice was performed randomly according to uniform distribution law. Other feature values in new sample were the same as in initial sample. Then unknown feature values were restored as the sample means.

Then they were «refined» with local method (method №1) or by solving the tasks of recognition (method №3). Optimization method was not considered, some preliminary results of its application are considered in [Ryazanov, 2011]. As a criterion for the solution of the restoration task were used two approaches: calculation of the quantity $\Psi_1(\alpha) = \sum_{\langle i,j\rangle \in J} \left|x_{ij} - x^*_{ij}\right| / |J|$ or $\Psi_2(\alpha) = \sqrt{\sum_{\langle i,j\rangle \in J} (x_{ij} - x^*_{ij})^2} / |J|$, where x^*_{ij} – restored feature values. The second approach is to measure the accuracy of the solution of the supervised classification problem on the recovered data.

4.1. Restoring gaps for model problem

The model problem has been formed as a mixture of random samples, obtained according to a normal distribution with 10 independent features. The advantage of the model problem is that we know the configuration of classes. Relatively «complicated» model example was considered. The first class is the union of four samples with 25 objects in each, features of which were created by normal distributions with means, respectively, 1, 5, 9, 13, and a variance 0.5. The second class was formed similarly, except that the features had mathematical expectations 3, 7, 11, 15. Visualization of the original sample of 200 objects is shown in Fig. 1. (rendering algorithm is described in [O Duda, Hart, 1973]). Figure 2 shows a visualization of a sample in which 50% of the random features in each object are replaced with sample mean for the sample. Figure 3 shows a visualization of the sample with missing data after unknown features values restoration. We see that the reconstructed data has a smaller spread in class than after replacing gaps with averages. In Tab. 1 the results of recognition accuracy under different levels of incompleteness α are shown. Each time a new sample was used. Implementation of well-

known methods was used in system RECOGNITION [Zhuravlev etc., 2006]: « k-nearest neighbors » (k-neighbors), « linear machine » (LM), « binary decision trees» (BT), «support vector machines » (SVM), «logical regularities» (LR) [Ryazanov, 2007]. During the experiments, the task to select control parameters of programs in order to build the best algorithms and further comparing algorithms hasn't been set. Values of their control parameters were usually chosen "by default". The aim was to compare different methods of reconstruction feature values and to study the behavior of the separate method depending on the level α. The recognition accuracy (the percentage of correctly recognized objects in the sliding mode control) of methods SVM and LM on the original samples has amounted, respectively, 75.0% and 54.5%. Therefore, on this problem were considered only « k- neighbors » methods (100%), «BT» (99.5%), «LR» (96.0%). Note that the number of nearest neighbors of the "default" is 3, which explains the high accuracy of the method in this example.

Figure1. Initial sample

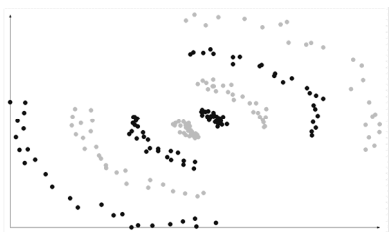


Figure 2. Replacement gaps with feature averages

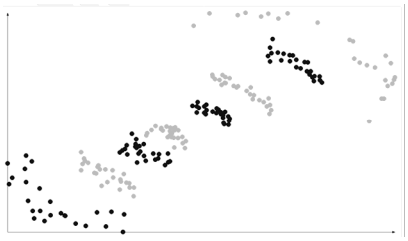


Figure 3. Restored data

method\α	10			20			30			40		
	avr.	loc.	recog.	avr.	loc.	recog.	avr.	loc.	recog.	avr.	loc.	recog.
κ-neighbors	90.0	100	100	94.0	100	100	82.0	97.0	100	70.5	98.0	99.5
BT	94.0	99.0	92.5	93.0	98.5	88.0	91.5	97.5	94.0	84.5	97.0	90.0
LR	72.5	98.0	86.5	50.5	97.0	81.0	52.5	95.0	81.0	51.5	94.5	85.5

method\α	50			60			70			80		
	avr.	loc.	recog.	avr.	loc.	recog.	avr.	loc.	recog.	avr.	loc.	recog.
κ-neighbors	67.5	97.0	97.0	64.5	90.5	84.5	66.5	87.0	66.5	63.5	78.0	72.8
BT	90.0	94.0	89.5	84.5	88.5	88.0	79.0	78.0	83.0	80.0	71.0	73.5
LR	58.5	86.5	85.0	47.0	80.0	74.0	44.0	64.0	64.0	55.5	66.5	70.0

Table 1. Recognition accuracy after the restoration of feature values

The use of recovery methods 1) and 3) significantly improves the accuracy comparatively to filling with averages. In gray are marked only those rare results when accuracy decreases. We see that for small and medium α the best accuracy is obtained by the "k-neighbors", which is not surprising due to compactness of the subclasses. The method of "LR" shows significantly higher results after application of 1) and 2) because the compactness of classes is considerably improved.

Mean values for the three methods of recognition accuracy are shown in Figure 4.

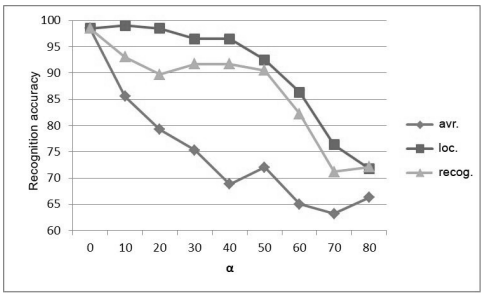


Figure 4. The average recognition accuracy on methods

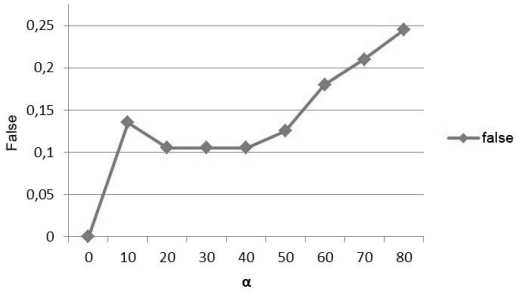


Figure5. Dependence the criterion $\Psi_2(\alpha)$ at model data

During replacing the gaps with mean values occurs that $\Psi_2(\alpha) \approx 0.26$. Dependence of the average values for the algorithm $\Psi_2(\alpha)$ is shown in Fig. 5. On other tasks, this relationship had a similar appearance.

4.2. Restoration of gaps on the example of "wine" task

Original data of this and future applications are taken from [Frank, Asuncion, 2010]. This problem was set by Forina, M. et al. (PARVUS - An Extendible Package for Data Exploration, Classification and Correlation. Institute of Pharmaceutical and Food Analysis and Technologies, Via Brigata Salerno, 16147 Genoa, Italy. We study a sample of 178 objects, which are the results of chemical analysis, expressed in 13 features, such as content of alcohol, malic acid, magnesium, hue and others. The sample is divided into three classes (59, 71 and 48 objects), corresponding to 3 types of wine made from grapes grown in the same region of Italy.

Visualization sample is shown in Fig. 6-8 is similar to visualization of model problem.

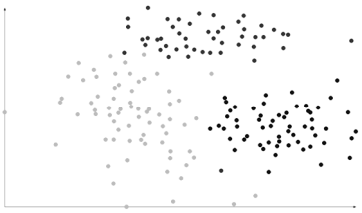


Figure6. Initial sample

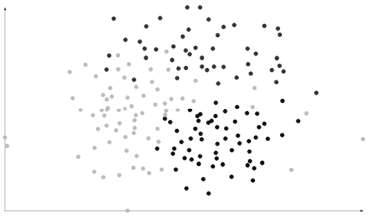


Figure 7. Replacement gaps with feature averages

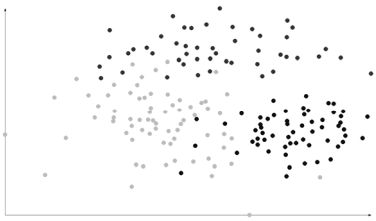


Figure 8. Restored data

In Tab. 2 the recognition accuracy under different levels of incompleteness α is shown.

метод\α	0	25			50			75		
		avr.	loc.	recog.	avr.	loc.	recog.	avr.	loc.	recog.
k-neighb.	97.2	84.8	93.8	96.1	83.1	83.1	89.3	76.4	75.8	79.2
LR	97.8	95.5	96.6	94.4	82.0	86.0	86.0	83.1	79.2	78.7
SVM	97.8	94.9	97.8	96.6	89.3	91.6	92.1	87.1	79.8	76.4
BT	94.9	86.0	92.7	93.8	62.9	77.0	86.0	76.4	70.2	68.5
LR	94.9	86.5	96.6	92.7	69.1	86.0	88.2	0.00	79.8	66.3

Table 2.

It can be seen that for small and medium α precision of recovery with methods 1) and 3) for subsequent recognition tasks is better than with the averaging method. However, with high level of gaps (here is the case of 75%) averages for the features are, apparently, more natural. The average value of accuracy here lowers the presence of LR method. During recovering feature values by the method of «mean» with a large number of gaps logical regularities in general are not found. Meanwhile, the average recognition accuracy without LR method is 80.5%.

4.3. Restoration of gaps on the example of "home" task

This problem of recognition the value of housing in the suburbs of Boston on set of 13 numerical signs (the crime rate in the city, the concentration of nitrogen oxides, the average number of rooms in the home, the age of the home, the weighted distance to five employment centers, etc.) is actually the problem of restoration of regression [Harrison, Rubinfeld, 1978]. The problem of classification with 5 classes was considered by dividing the cost of the house at intervals. Each class included, respectively, 16, 91, 93, 27 and 15 objects. The specifics of this problem was in "ordering of classes." The accuracy of recognition by the five methods is presented in Tab. 3:

metho d\α	0	15			30			40			50		
		avr.	loc.	recog.	avr.	loc.	recog.	avr.	loc.	recog.	avr.	loc.	recog.
κ-nb.	66.9	62.8	66.5	57.4	57.0	55.8	56.2	49.0	58.3	41.7	51.2	42.6	39.7
LR	73.6	68.6	66.9	67.4	65.7	65.7	67.4	62.8	58.7	57.9	59.9	48.3	55.0
SVM	74.0	67.8	70.7	70.7	66.5	67.8	64.9	54.1	61.2	55.4	54.9	57.9	61.6
BT	65.3	65.7	64.9	59.5	53.7	61.6	64.5	55.8	53.3	61.2	50.0	54.1	47.1
LR	64.5	62.0	63.6	62.8	61.2	54.5	58.3	49.2	55.4	52.1	50.4	54.5	53.3

Table 3.

Here, the quality of all methods is about the same, due, apparently, to simplicity of the task.

4.4. Restoration of gaps on the example of "image" task

The problem of classification 7 types of images ("heaven," grass "," cement ", etc.) was considered. Data source: Vision Group, University of Massachusetts, Carla Brodley. Image Segmentation data. Each class was represented by 30 precedents, each of which was described by a set of 19 numerical features. The accuracy of recognition with five methods is shown in Tab 4, and the average accuracy graph (of recognition methods) is shown in Fig. 9:

method\α	0	10			25			50		
		avr.	loc.	recog.	avr.	loc.	recog.	avr.	loc.	recog.
κ-nb.	87.6	81.4	83.3	77.6	73.8	79.5	82.4	63.3	68.6	66.8
LR	88.1	81.9	83.3	82.4	76.2	82.4	75.4	60.3	68.9	62.7
SVM	90.0	83.8	86.2	83.8	76.2	83.8	83.8	63.5	70.6	69.0
BT	84.8	85.2	84.8	82.9	70.5	82.4	76.2	59.4	69.0	59.7
LR	87.6	82.9	89.0	85.2	75.2	85.2	80.0	29.5	68.6	63.3

Table 4.

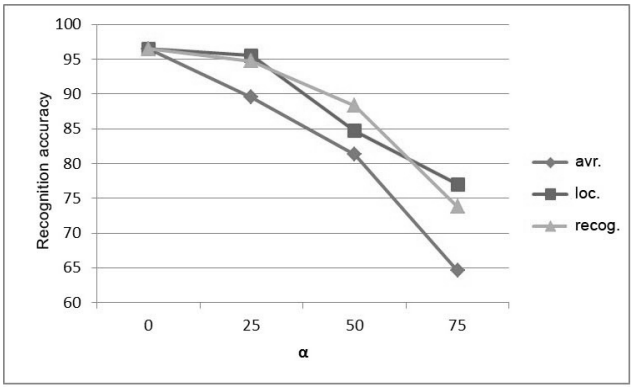


Figure 9

Conclusion

Experiments conducted confirm a priori expectations. The accuracy of reconstruction of feature values with methods 1, 3, is generally higher than simple averaging features. This is confirmed by both criteria: higher values of the classification accuracy, and lower values of the mean square deviation of feature values. The ratio of number of tasks where the classification accuracy of the solution after application of 1st and 3rd method was higher than accuracy of solving similar problems after averaging feature values is 107:37. Note that the thesis "the more gaps the worse recognition" is sometimes not fulfilled, due to small sample sizes and conducting of each experiment on a new sample. Creation of new algorithms to recover unknown feature values is important.

Having a set of N different recovery algorithms $A^i, i = 1,2,...,N$, we can build sets $\{\mathbf{x}^i, i = 1,2,...,N\}$ of complete descriptions, corresponding to the original incomplete description $\mathbf{x}$. Then, solving the problem of classification for each of the objects $\{\mathbf{x}^i, i = 1,2,...,N\}$ and using the procedure of constructing collective decisions, we can construct the final classification for $\mathbf{x}$. It appears that the collective classification accuracy of $\mathbf{x}$ is higher than the accuracy of the classification of individual $\{\mathbf{x}^i, i = 1,2,...,N\}$.

This work was supported by the Program of the Presidium of RAS № 15 and № 5 "Basic Sciences - Medicine", program №2 of Department of Mathematical Sciences RAS, RFBR projects 12-01-00912, 12-01-90012 Bel_a, 11-01-00585.

Bibliography

[Delavallade, Dang, 2007] T. Delavallade and T.H. Dang. Using Entropy to Impute Missing Data in a Classification Task. In: IEEE International Conference on Fuzzy Systems, London, 1-6, 2007.

[Frank, Asuncion, 2010] A. Frank and A. Asuncion. UCI Repository of machine learning databases. In: http://www.ics.uci.edu/~mlearn/MLRepository.html. Irvine, CA: University of California, Department of Information and Computer Science, 2010.

[Harrison, Rubinfeld, 1978]Harrison, D. and Rubinfeld, D.L. Hedonic prices and the demand for clean air, J. Environ. In: Economics & Management, vol.5, 81-102, 1978.

[Honghaj etc., 2005] F. Honghai, C. Guoshun, Y. Cheng, Y. Bingru and C. Yumei. A SVM Regression Based Approach to Filling in Missing Values. In: LNCS - Knowledge-Based Intelligent Information and Engineering Systems, Springer Berlin - Heidelberg, vol. 3683, 581-587, 2005.

[Little, Rubin, 1987] R.J.A. Little, D.B. Rubin. Statistical Analysis with Missing Data. In: Wiley, New York, 1987.

[Morin, Raeside, 1981] R. L. Morin and D. E. Raeside. A reappraisal of distance-weighted k-nearest neighbor classification for pattern recognition with missing data. In: IEEE Transactions on Systems, Man and Cybernetics 11(3): 241–243, 1981

[O Duda, Hart, 1973] Richard O.Duda, Peter E.Hart. Pattern Classification and Scene Analysis. In: A Wiley-Interdcience Publication John Wiley&Soms, 1973.

[Ryazanov, 2007] V.V. Ryazanov. Logical Regularities in Pattern Recognition (Parametric Approach). In: Computational Mathematics and Mathematical Physics, Vol. 47, No. 10, pp. 1720–1735. © Pleiades Publishing, Ltd., 2007. Original Russian Text © V.V. Ryazanov, published in Zhurnal Vychislitel'noi Matematiki i Matematicheskoi Fiziki, 2007, Vol. 47, No. 10, pp. 1793–1808, 2007

[Ryazanov, 2011] V.V. Ryazanov. Some Imputation Algorithms for Restoration of Missing Data. In: "Lecture Notes in Computer Science" (LNCS), vol. 7042, pp. 372-379, 2011.

[Zhang, 2008] S. Zhang. Parimputation: From imputation and null-imputation to partially imputation.In: IEEE Intelligent Informatics Bulletin, vol. 9(1), 32-38, 2008.

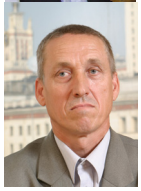
[Zhuravlev etc., 2006] Yu. I. Zhuravlev, V.V. Ryazanov, O.V. Senko. Recognition. Mathematical methods. The software system. Applications. In: Moscow: Fazis, 2006.

[Zhuravlev, Nikiforov, 1971] Yu. I. Zhuravlev and V. V. Nikiforov. "Recognition Algorithms based on Estimate Evaluation". In: Kibernetika, No. 3, pp. 1–11, 1971.

[Zloba, 2002] E. Zloba. Statistical methods of reproducing of missing data. In: Computer Modelling & New Technologies. —. — Vol.6, No.1 — P. 51-61, 2002

Authors' Information

Vasily V. Ryazanov *– Student; Moscow Institute of Physics and Technologies, Russia, 141700, Moscow reg., Dolgoprudny, Institutsky per. 9; e-mail: vasyarv@mail.ru*
Major Fields of Scientific Research: Data Mining, Missing Data, Mathematical models in Economics.



Vladimir V. Ryazanov *– Heard of Department; Institution of Russian Academy of Sciences Dorodnicyn Computing Centre of RAS, Russia,* 119991, Moscow, Vavilov's street, 40; *e-mail: rvv@ccas.ru*
Major Fields of Scientific Research: Pattern recognition, Data mining, Artificial Intelligence