Dynamic Clustering-Based Estimation of Missing Values in Mixed Type Data*

Vadim Ayuyev[1], Joseph Jupin[2], Philip W. Harris[3] and Zoran Obradovic[2]

[1] FN1-KF Department, Bauman Moscow State Technical University (Kaluga Branch), Bazgenova Str. 2, Kaluga, 248600, Russian Federation
vadim.ayuyev@gmail.com
[2] Center for Information Science and Technology, Temple University, 303 Wachman Hall, 1805 N. Broad St., Philadelphia, PA 19122, USA
joejupin@temple.edu, zoran@ist.temple.edu
[3] Department of Criminal Justice, Temple University, 512 Glatfelter Hall, 1115 W Berks Str., Philadelphia, PA 19122, USA
Phil.Harris@temple.edu

Abstract. The appropriate choice of a method for imputation of missing data becomes especially important when the fraction of missing values is large and the data are of mixed type. The proposed dynamic clustering imputation (DCI) algorithm relies on similarity information from shared neighbors, where mixed type variables are considered together. When evaluated on a public social science dataset of 46,043 mixed type instances with up to 33% missing values, DCI resulted in more than 20% improved imputation accuracy over Multiple Imputation, Predictive Mean Matching, Linear and Multilevel Regression, and Mean Mode Replacement methods. Data imputed by 6 methods were used for test of NB-Tree, Random Subset Selection and Neural Network-based classification models. In our experiments classification accuracy obtained using DCI-preprocessed data was a lot better than when relying on alternative imputation methods for data preprocessing.

Keywords: Data pre-processing, data imputation, clustering, classification.

1 Introduction

A common approach to analyzing data with missing values is to remove attributes and/or instances with a large fraction of missing values. Such data preprocessing is appealing because it is simple and also reduces dimensionality. However, this is not applicable when missing values cover a lot of instances, or their presence in essential attributes is large [1].

Another common and practical way to address the problem of missing values in data is to replace them as estimates derived from the non-missing values by a linear function. The missing attribute j from an instance i, denoted as $x_{i,j}^{ms}$, is estimated as:

* Contact: Z. Obradovic, phone: +1.215.204.6265, fax: +1.215.204.5082, alternative e-mail: zoran.obradovic@temple.edu

$$x_{i,j}^{ms} = f\left(x_{1,j}, x_{2,j}, \ldots, x_{p,j}, \ldots, x_{P_j,j}\right), \quad (1)$$

where f is a linear function of P_j variables; P_j is the number of instances in the data with non-missing values for attribute j; and $x_{p,j}$ is a non-missing attribute j from an instance p.

A special case of (1), which is simple, fast, and often provides satisfactory results when the number of missing values is relatively small and their distribution is random, is mean (or mode for categorical attributes) value based imputation:

$$x_{i,j}^{ms} = \frac{1}{P_j}\sum_{p=1}^{P_j} x_{p,j} \quad . \quad (2)$$

The limitation of mean value based imputation and its variations is its focus on a specific variable without taking into account the overall similarities between instances. For example, consider the following 5 data points with 6 attributes, where a categorical attribute (fifth column) is missing one value (denoted as "*ms*"):

$$\begin{bmatrix} 1 & 10.2 & 1 & 1 & ms & 1 \\ 1 & 9.8 & 1 & 1 & 2 & 1 \\ 0 & 1.1 & 0 & 0 & 1 & 0 \\ 0 & 1.1 & 0 & 0 & 1 & 1 \\ 1 & 0.3 & 0 & 0 & 1 & 0 \end{bmatrix}. \quad (3)$$

Here, it would be reasonable to replace "*ms*" by "2" since the first two instances are very similar. However, mean/mode value-based imputation methods would replace "ms" by "1" as it is the most common value for this attribute in the dataset.

One of the most powerful approaches to missing values estimation is replacement by multiple imputation [1, 2]. The idea is to generate multiple simulated values for each incomplete instance, and iteratively analyze datasets with each simulated value substituted in turn. The purpose is to obtain estimates that better reflect the true variability and uncertainty in the data than are done by regression. Multiple imputation methods yield multiple imputed replicate datasets each of which is analyzed in turn. The results are combined and the average is reported as the estimate. For continuous attributes and a fairly small fraction of missing values, reliable estimates are obtained by combining only a few imputed datasets.

A clustering based approach for missing data imputation was considered as a local alternative to global estimation [3]. The premise was that instances could be grouped such that all the imputations in identified groups are independent from other groups. However, previous distance-based [4] clustering work was focused mainly to development of supervised clustering methods and mean/mode based imputations in these clusters. Also, prior studies were based on a strict separation for objects within clusters, such that it was assumed that there is no influence of instances in one cluster to an imputation process in other clusters.

In our DCI approach an independent cluster of similar instances with no missing values for a particular attribute is constructed deterministically around each instance with a missing value. In contrast to a typical clustering method, we allow cluster intersections such that the same instance may be included in many clusters. DCI relies on a distance measure that considers together categorical and continuous variables and is applicable for estimation of missing values in high dimensional mixed type data.

2 Methodology

We assume that the given data consist of M instances with N attributes where N is a mixture of tens to hundreds of categorical and continuous attributes. For the proposed Dynamic Clustering based Imputation (DCI) we use a dissimilarity measure between instances in a mixed type dataset described in Section 2.1. This measure is used in a clustering algorithm for identification of similar instances as described in Section 2.2 to perform a dynamic cluster-specific imputation of missing values as described in Section 2.3. An evaluation method and alternative imputation approaches were described in Section 2.4.

2.1 Measuring Dissimilarity between Instances in Mixed Type Data for DCI

The Minkowski distance, the Simple Matching Coefficient, the Jacquard Similarity Coefficient or other metrics could be used separately to measure the distance between instances for each type of attributes. However, such approaches are of limited applicability for mixed type data consisting of categorical and continuous attributes in the presence of many missing values [5]. In DCI given N dimensional data, to measure the dissimilarity between two instances x_i and x_j of mixed type in the presence of missing values, we compute [6]:

$$\mathrm{dst}\left(x_i, x_j\right) = \left[\sum_{n=1}^{N} \frac{\left|x_{i,n} - x_{j,n}\right|}{\max\limits_{p=1\ldots P_n} x_{p,n} - \min\limits_{p=1\ldots P_n} x_{p,n}} \delta_{i,j}^{(n)}\right] \Bigg/ \sum_{n=1}^{N} \delta_{i,j}^{(n)} \quad , \qquad (4)$$

$$\delta_{i,j}^{(n)} = \begin{cases} 0, & \text{if one of } x_{i,n} \text{ or } x_{j,n} \text{ is missing;} \\ 1, & \text{otherwise} \end{cases}$$

where max and min are computed over all non-missing vales of the n-th attribute.

2.2 Clustering for Identification of Similar Instances in DCI

To identify similar instances in DCI we employ a new clustering algorithm consists of the following steps:

1. Computing the similarity matrix (*SM*) for all instances:

$$SM = \begin{pmatrix} \infty & \mathrm{dst}(x_1, x_2) & \dots & \mathrm{dst}(x_1, x_M) \\ \mathrm{dst}(x_2, x_1) & \infty & \dots & \mathrm{dst}(x_2, x_M) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{dst}(x_M, x_1) & \mathrm{dst}(x_M, x_2) & \dots & \infty \end{pmatrix}. \quad (5)$$

2. Computing the neighborhood matrix (*NM*):

$$NM = \begin{pmatrix} nm_{1,1} & \dots & nm_{1,M} \\ \vdots & \ddots & \vdots \\ nm_{M,1} & \cdots & nm_{M,M} \end{pmatrix}, \quad (6)$$

where $nm_{i,j}$ is the number of common neighbors among K nearest neighbors for instances i and j, and M is the total number of instances in the dataset.

3. Constructing an ordered list (by ascending sort) of all neighbor instances with no missing value in j-th attribute for each missing value $x_{i,j}^{ms}$:

$$list_{i,j} = \mathrm{sort}\left\{ \mathrm{dst}\left(x_i^{ms}, x_p\right) / nm_{i,p}, \quad p = \overline{1, P_j}; nm_{i,p} > 0 \right\}, \quad (7)$$

where x_i^{ms} denotes i-th instance with missing value in j-th attribute, and x_p denotes p-th instance with no missing in j-th attribute. Here, if two instances have the same dst/*nm* rate, one with less missing attributes is listed first in the list.

4. Constructing a cluster $C_{i,j}$ for each missing value $x_{i,j}^{ms}$ by using first R elements of $list_{i,j}$, where R is a user-specific parameter that defines a cluster size, and $R < |list_{i,j}|$.

2.3 Dynamic Cluster-Specific Imputation Methods for Mixed Type Data

In a cluster constructed as described in Section 2.2 using similarity measure introduced in Section 2.1 a missing value could be imputed based on (a) the mean value of the corresponding attribute in other items contained in this cluster, or (b) similarity to the nearest instance with a non-missing value. Averaging in (a) and identification of the nearest instances from the same cluster in (b) could be based on various metrics. In DCI, we use the following categorical and continuous data specific metrics aimed to provide a balance in terms of imputation quality and computational complexity:

Categorical variable: A missing value is estimated by the corresponding attribute in an instance from the same cluster that has the largest number of common neighbors with the imputed instance:

$$x_{i,j}^{ms} = x_{r,j}^{(C_{i,j})} : \max_{r=1\ldots R}\left\{nm_{i,r}\right\} \ . \tag{8}$$

Continuous variable: A missing value is estimated based on all instances in the same cluster where each non-missing value is weighted by the appropriate entry of the neighborhood matrix *NM*:

$$x_{i,j}^{ms} = \left[\sum_{r=1}^{R} x_{r,j}^{(C_{i,j})} nm_{i,r}\right] \Big/ \sum_{r=1}^{R} nm_{i,r} \ . \tag{9}$$

2.4 Evaluation Measures and Alternative Imputation Methods

For evaluating imputation quality different measures were used when comparing imputed categorical and imputed numerical data versus the corresponding true values.

The mean and absolute squared error measurements tend to be very sensitive to outliers. So, for continuous attributes and for a given tolerance τ we measured a Relative Imputation Accuracy (*RIA*, also known as relative prediction accuracy [7]) defined as

$$RIA_{\tau} = \frac{n_{\tau}}{Q} \times 100\% \ , \tag{10}$$

where n_{τ} is the number of imputed elements estimated within τ percent of accuracy from the true value of the corresponding missing value and Q is the total number of imputed values in the data. *RIA* is very useful in practice as an absolute precision of imputed continuous values is often not needed. A nice property of *RIA* measure is that it is not affected by an individual incorrect imputation (e.g. large value instead of small) that could affect considerably some statistical measures (e.g., *MSE*-based [8]).

In categorical attributes we measured a fraction of Correct Imputations (*CI*) defined as

$$CI = \frac{s}{Q} \times 100\% \ , \tag{11}$$

where s is the number of correctly estimated imputed elements.

As a simple imputation alternative to DCI, we used a WEKA implementation [9] of Mean and Mode Replacement (denoted here as MMR). We also compared DCI to four statistically well-founded techniques: Multiple Imputation [1, 2], Predictive Mean Matching [10] (denoted here as PMM), Linear Regression [11], and Multilevel Regression [11] (denoted here as MLR).

The Multiple Imputation Method used for comparison and implemented in Amelia II software [12] enables to draw random simulations from the multivariate normal observed data posterior, and uses standard Expectation Maximization (EM) for finding an appropriate set of starting values for data argumentation. Multiple Imputation begins with EM and adds an estimation of uncertainty for receiving draws from the correct posterior distribution followed by a resampling based on importance. According to King et all [12], this way is faster than traditional multiple imputation approaches, does not rely on Markov chains, produces the fully independent imputations and allows the use of about 50% more information.

Predictive Mean Matching comparison method implemented in WinMICE software [13] combines both parametric and nonparametric techniques. It imputes missing values by means of the nearest neighbors where the distance is computed as the expected values of the missing variables conditional on the observed covariates, instead of directly on the values of the covariates.

Linear and Multilevel Regression models, also implemented in WinMICE, are well known statistical approaches that allow variance in imputed variables to be analyzed at multiple hierarchical levels, whereas in linear regression all effects are modeled to occur at a single level.

3 Results and Discussion

We first performed experiments on a social science dataset with mixed-type attributes to compare quality of imputation by the proposed method and alternatives in presence of various fractions of missing values. In another set of experiments mixed-type data preprocessed by various imputation methods was used for classification by several algorithms to determine practical effects of an imputation method on classification accuracy (reported in Section 3.2).

A public domain *Adult* dataset [14], from the UCI Machine Learning Repository was used for comparing different data imputation methods. The dataset contained a subset of records about the US population collected by the US Census Bureau. The 48,842 individuals in this database are described by 8 categorical and 6 continuous attributes (with some missing data) related to prediction of annual income. In our experiments etalon data with 46,043 instances was constructed by removing all instances from the *Adult* dataset with missing values. To make the dataset balanced in terms of different attribute types, two categorical attributes ("education" and "native country") were also removed.

Eight test datasets with missing values ("holes") were constructed by randomly hiding 0.2%, 0.5%, 1.1%, 1.8%, 5.4%, 10.9%, 16.3% and 32.6% of data elements in both categorical and continuous attributes of the etalon data. Each test database was fully independent from others, which means that places of "holes" were independent.

3.1 Evaluation of Imputation Quality on Mixed Type Data

The DCI and other imputation algorithms described in section 2.4 were compared using eight datasets with different fraction of introduced missing values. Imputed

values were compared to the true values in *Adult* dataset. As to provide a comparison to a trivial estimate, we also report the results obtained by using the corresponding attribute value in one of randomly selected instances (denoted here as Random). The imputation accuracy by DCI and alternative methods are summarized in Tables 1-4. All reported DCI results were obtained for 50 nearest neighbors and 9 the most common neighbors (K=50, R=9 defined in section 2.2). Very similar findings (within 10% of reported) were obtained for 40<K<60 and R=7 or R=11, but were less robust for K<20 or R>15 (stability results are omitted due to lack of space).

Table 1. Fraction of correct imputation (*CI*) in categorical attributes for 0.2%-32.6% imputed values.

Imputation Methods	*CI* for different fractions of missing values							
	0.2%	0.5%	1.1%	1.8%	5.4%	10.9%	16.3%	32.6%
DCI	66.0	69.2	67.5	70.0	71.0	71.5	70.6	65.6
MMR	54.5	38.0	53.7	56.2	54.4	54.7	54.7	54.7
PMM	34.2	37.8	35.9	36.1	36.8	35.9	35.6	35.2
Linear Regression	28.1	30.3	28.8	29.0	29.0	28.4	28.4	28.1
Multiple Imputation	46.9	49.1	49.0	49.8	48.1	47.8	47.3	45.5
MLR	29.3	29.7	27.9	29.8	28.8	28.5	28.6	28.2
Random	19.1	20.8	19.3	21.1	19.9	20.3	19.7	20.2

Imputation accuracy results for estimation of categorical attributes (Table 1) revealed that for all fractions of missing values DCI was a lot more accurate from alternative five imputation methods (1.2-1.4 times more accurate than the best of the remaining methods). Mean Mode Replacement approach was the second most accurate imputation method for categorical attributes. The results of the remaining imputation methods had more than 50% imputation error, but were still much better than random replacements.

The Relative Imputation Accuracy of DCI for imputation of continuous attributes (Tables 2-4) was also much better from alternative imputation methods. Here, Predictive Mean Matching was the second most accurate method. For 5% tolerance DCI provided 1.4-1.8 times better accuracy than PMM and was 6-9 times other better than alternatives (Table 2). Still, even the weak imputation methods were significantly more accurate than random replacements.

Table 2. Relative imputation accuracy (*RIA*) with 5% tolerance in continuous attributes for 0.2%-32.6% imputed values.

Imputation Methods	*RIA* (τ = 5%) for different fractions of missing values							
	0.2%	0.5%	1.1%	1.8%	5.4%	10.9%	16.3%	32.6%
DCI	33.8	28.3	28.1	31.2	29.5	30.3	30.2	28.3
MMR	3.7	4.9	1.4	5.5	1.2	1.5	1.4	5.5
PMM	18.6	20.9	20.0	20.2	18.7	19.6	19.4	19.4
Linear Regression	3.7	4.5	4.4	4.5	4.2	4.3	4.4	4.2
Multiple Imputation	5.5	11.8	3.9	4.7	4.6	4.7	4.6	4.5
MLR	3.7	4.3	4.6	4.0	4.2	4.4	4.4	4.3
Random	1.8	2.1	2.9	3.3	3.0	3.2	3.0	3.0

Table 3. Relative imputation accuracy (*RIA*) with 10% tolerance in continuous attributes for 0.2%-32.6% imputed values.

Imputation Methods	*RIA* (τ= 10%) for different fractions of missing values							
	0.2%	0.5%	1.1%	1.8%	5.4%	10.9%	16.3%	32.6%
DCI	38.7	35.4	35.6	37.4	36.7	37.2	37.2	35.6
MMR	10.2	11.8	12.0	11.9	11.5	11.6	11.6	12.0
PMM	25.6	30.0	29.2	28.8	27.6	28.5	28.2	28.2
Linear Regression	10.7	13.6	13.6	13.0	13.1	13.3	13.2	13.1
Multiple Imputation	13.3	20.4	13.4	13.5	13.3	13.3	13.4	13.3
MLR	10.7	12.9	13.9	13.0	13.1	13.2	13.3	13.2
Random	10.5	12.2	13.2	12.8	12.8	13.0	12.9	12.8

Table 4. Relative imputation accuracy (*RIA*) within 15% tolerance in continuous attributes for 0.2%-32.6% imputed values.

Imputation Methods	*RIA* (τ= 15%) for different fractions of missing values							
	0.2%	0.5%	1.1%	1.8%	5.4%	10.9%	16.3%	32.6%
DCI	42.0	40.1	40.0	41.8	41.5	42.0	42.0	40.5
MMR	15.6	17.6	17.5	17.4	17.4	17.4	17.4	17.8
PMM	30.3	34.7	33.9	33.5	32.6	33.5	33.1	33.3
Linear Regression	15.4	18.3	18.4	17.7	18.1	18.3	18.2	18.1
Multiple Imputation	17.0	25.5	18.1	17.8	18.2	18.1	18.2	18.1
MLR	15.0	18.1	18.5	17.5	18.2	18.2	18.2	18.2
Random	15.8	17.7	18.0	17.9	18.2	18.3	18.2	18.2

Experiments with double and triple tolerance for allowed estimation error of 10% and 15% (Tables 3 and 4) resulted in reduced differences in accuracy between imputation methods. However, even for larger tolerance DCI was still 20-50% more accurate (in relative difference) than the second best PMM method. These experiments suggest that the Mean Mode Replacement, Regression methods, and even Multiple Imputation methods are not appropriate for larger tolerance estimation in continuous variables as the corresponding results were comparable to random replacement. On the other hand, all methods outperformed the Mean Mode Replacement, which is commonly used in practice due to its simplicity.

3.2 Effect of an Imputation Method on Classification Accuracy for Mixed Type Data

The next stage of our experiments was devoted to practical comparison of how well different imputation techniques would suit for real life classification tasks. The idea was to explore a scenario where clean mixed type data was used for training a classification model while it was applied to real data with various fractions of missing values. For this purpose we built several kinds of classifiers by training them on the first 16,043 subjects from etalon *Adult* database where for each instance all 12 attributes were available. For a test subject drawn from the remaining 30,000 instances the task was to predict if he/she makes over 50,000 U.S. dollars a year

where a fraction of variables was missing at random. Different fractions of missing values were considered and preprocessing was achieved by 6 imputation methods described in Section 2. As a measure of accuracy, the percent of correctly classified instances was calculated.

As a classification method we applied three models implemented in WEKA: NB-Tree [15], Random Subset Selection [16] and Multilayer Perceptron [17]. NB-Tree was used as one of the best classification methods for *Adult* database according to [14]. Random Subset Selection and Multilayer Perceptron were used as alternative solutions that in other domains showed good speed and classification accuracy, respectively. The classification results reported in Tables 5-7 are compared to the upper bound obtained by testing on complete data without missing values.

Table 5. Classification Accuracy (*CA*) of NB-Tree classification model applied to datasets with 0.2%-32.6% imputed values.

Imputation Methods	*CA* of NB-Tree for different fractions of missing values							
	0.2%	0.5%	1.1%	1.8%	5.4%	10.9%	16.3%	32.6%
DCI	86.1	86.1	86.0	86.0	85.8	85.4	84.8	83.6
MMR	86.1	86.0	85.9	85.9	85.1	84.2	83.0	79.6
PMM	86.1	86.1	85.9	85.9	85.0	84.5	83.6	81.0
Linear Regression	86.1	86.0	85.7	85.6	84.3	82.9	81.4	76.8
Multiple Imputation	86.1	86.1	85.7	85.7	84.6	83.1	81.4	77.1
MLR	86.1	86.0	85.8	85.6	84.3	82.9	81.4	76.8
Upper bound	86.1	86.1	86.1	86.1	86.1	86.1	86.1	86.1

All imputation methods resulted in very similar accuracy for small fractions (0.2-1.8%) of missing values (Table 5). However, the difference was substantial when more than 10% of missing values were imputed. Though DCI provided the most accurate NB-Tree classifier for all fractions of missing values, its advantage was the most evident for the largest fraction of missing values (32.6%) where it had 14-22% less relative difference in error than alternative methods.

Table 6. Classification Accuracy (*CA*) of Random Subspace Selection classification model applied to datasets with 0.2%-32.6% imputed values.

Imputation Methods	*CA* of Random Subset for different fractions of missing values							
	0.2%	0.5%	1.1%	1.8%	5.4%	10.9%	16.3%	32.6%
DCI	84.9	84.9	84.9	84.8	84.8	85.0	84.7	84.8
MMR	84.9	84.9	84.9	84.8	84.5	84.3	83.8	81.6
PMM	84.9	84.9	84.8	84.7	84.4	84.2	84.1	83.1
Linear Regression	84.8	84.9	84.8	84.7	84.2	83.7	83.4	82.1
Multiple Imputation	84.9	84.9	84.8	84.8	84.5	84.2	83.7	82.6
MLR	84.9	84.9	84.8	84.7	84.3	83.8	83.3	81.8
Upper bound	84.9	84.9	84.9	84.9	84.9	84.9	84.9	84.9

When using Random Subset Selection classifier the overall results were consistent to classification by NB-Tree classifier (Table 6). However, Random Subset Selection classifier was more tolerant to increase in fraction of missing values. Once again, DCI

outperformed alternative approaches on the largest fractions of missing values for an 11-20% relative difference in error.

Neural Network based classifier, represented by a 3-layer Perceptron showed similar characteristics to NB-Tree and Random Subspace Selection (Table 7). DCI imputation resulted in more accurate classification in all datasets with a large fraction of missing values. For 0.5%, 1.8%, 5.4% 10.9%, and 32.6% imputed values a neural network achieved somewhat better accuracy than the upper bound obtained on etalon data without missing values. This may be accounted to noise tolerant property of a Multilayer Perceptron.

Table 7. Classification Accuracy (*CA*) of Multilayer Perceptron classification model applied to datasets with 0.2%-32.6% imputed values.

Imputation Methods	*CA* of Multilayer Perceptron for different fractions of missing values							
	0.2%	0.5%	1.1%	1.8%	5.4%	10.9%	16.3%	32.6%
DCI	84.5	84.6	84.5	84.6	84.6	84.7	84.5	84.7
MMR	84.5	84.5	84.4	84.4	84.1	83.6	83.0	80.8
PMM	84.5	84.5	84.3	84.3	83.8	83.2	82.7	80.8
Linear Regression	84.5	84.5	84.2	84.1	83.2	82.0	81.0	77.4
Multiple Imputation	84.5	84.5	84.3	84.2	83.4	82.3	81.2	77.5
MLR	84.5	84.5	84.3	84.2	83.2	82.2	80.9	77.5
Upper bound	84.5	84.5	84.5	84.5	84.5	84.5	84.5	84.5

To address a considerable class misbalance for target variable in the *Adult* dataset (34,621 subjects in one class vs. 11,422 in another) we also measured Kappa coefficient [18] and F-score [19] for three classification models when imputing 32.6% of missing values by six methods (Table 8).

Table 8. Kappa coefficient and F-score of NB-Tree, Random Subspace Selection, and Multilayer Perceptron classification models applied to datasets with 32.6% of missing values imputed by 6 methods and to complete data without missing values.

Imputation Methods	NB-Tree		Random Subset		Multilayer Perceptron	
	κ	F	κ	F	κ	F
DCI	0.52	0.83	0.54	0.84	0.54	0.83
MMR	0.42	0.79	0.49	0.81	0.45	0.80
PMM	0.47	0.81	0.47	0.81	0.44	0.80
Linear Regression	0.37	0.77	0.44	0.80	0.37	0.77
Multiple Imputation	0.40	0.77	0.48	0.81	0.38	0.77
MLR	0.37	0.77	0.43	0.80	0.38	0.77
Upper bound	0.61	0.86	0.55	0.84	0.53	0.83

Here, Kappa coefficient is defined as:

$$\kappa = \frac{Ra - Pa}{1 - Pa}\ , \qquad (\mathbf{12})$$

where *Ra* is the relative observed agreement, and *Pa* is the hypothetical probability of chance agreement, using the observed data to calculate the probabilities of each classifier randomly choosing each category.

F-score is defined as:

$$F = 2\frac{precision \times recall}{precision + recall}\ , \qquad (13)$$

where, in a classification context, *precision* denotes the number of true positive predictions divided by the total number of items labeled as positive in the data set, while *recall* denotes the number of true positive predictions divided by the total number of items that were predicted as positive.

The obtained results clearly suggest that DCI based pre-processing results in the nearest accuracy to the upper bound in the sense of both Kappa coefficient and F-score statistics. We also observe that our results on imputed data confirms previous findings obtained on complete data that NB-Tree based classifier is a good choice for classification of *Adult* data. However, we also observe that the most stable results in terms of accuracy were obtained by Random Subset classifier.

4 Conclusion

Data imputation to replace missing values is often an important preprocessing step in data analysis. This study identified some limitation of a commonly used heuristic and of four known statistical methods when applied to mixed type data with a large fraction of missing values. In our approach, the main idea was to make all replacements independently for data within clusters created around each missing value. This is theoretically reasonable and is useful for a practical implementation. Our experiments on a social science mixed type data provide evidence that the proposed data imputation method is more accurate than the evaluated alternatives and is effective when a large fraction of data is missing.

While the computational complexity of the proposed imputation method of $O(M^3 \log M)$ could be a limiting factor in large scale applications, many possibilities for improvements remain. For example, cluster-specific multiple imputation techniques based on DCI idea could be developed. Also, specialized algorithms for defining optimal size of specific clusters may be created. Finally, organizing data to KD-trees may improve the overall matrix processing speed.

Acknowledgments. This study was funded in part by Award No. 2006-IJ-CX-0022 awarded by the National Institute of Justice, Office of Justice Programs, US Department of Justice. The opinions, findings and conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of the Department of Justice. Funding support from Russian Ministry of Science and Education is also acknowledged.

Authors thank the Department of Human Services (DHS), Philadelphia, and the Crime and Justice Research Center at Temple University for creating ProDES [20]

and granting us access to the dataset. During this study we found a real-life data imputation problem that initiated DCI algorithm development.

We also thank Dr. Alan Izenman and Dr. Jeremy Mennis for extremely valuable comments on an early version of the methods and the results reported in this manuscript. Finally, we thank Pavel Karpukhin who rewrote all Matlab code, which made it possible to do most of our experiments.

References

1. Little, R.J.A., Rubin, D.B.: Statistical analysis with missing data. John Wiley & Sons, New York (1987)
2. Schafer, J.L.: Multiple imputation: a primer. Statistical Methods in Medical Research. Vol. 8, 1, 3--15 (1999)
3. Fujikawa, Y., Ho, T.B.: Cluster-based algorithms for dealing with missing values. In: Knowledge Discovery and Data Mining Conference, pp. 549--554. Springer, Berlin (2002)
4. Mantaras, R.L.: A distance-based attribute selection measure for decision tree induction. Machine Learning. Vol. 6, 81--92 (1991)
5. Gan, G., Ma, C., Wu, J.: Data Clustering: Theory, Algorithms, and Applications. SIAM Press, Philadelphia (2007)
6. Wishart, D.: K-means clustering with outlier detection, mixed variables and missing values. In: Schwaiger M., Opitz O. (eds.) Exploratory Data Analysis in Empirical Research, pp. 216--226. Springer, New York (2003)
7. Nelwamondo, F.V., Mohamed, S., Marwala, T.: Missing Data: A comparison of neural network and expectation maximization techniques. Current Science. Vol. 3, 11, 1514--1521 (2007)
8. Bermejo, S., Cabestany, J.: Oriented principal component analysis for large margin classifiers. Neural Networks. Vol. 14, 10, 1447--1461 (2001)
9. Witten, I.H., Frank, E.: Data Mining: Practical machine learning tools and techniques, 2nd Edition. Morgan Kaufmann, San Francisco (2005)
10. Landerman, L.R., Land, K.C., Pieper, C.F.: An Empirical Evaluation of the Predictive Mean Matching Method for Imputing Missing Values. Sociological Methods & Research, Vol. 26, 1, 3--33 (1997)
11. Gelman, A., Hill, J.: Data Analysis Using Regression and Multilevel/Hierarchical Models. Cambridge University Press, Cambridge (2006)
12. King, G., Honaker, J., Joseph, A., Scheve, K.: Analyzing Incomplete Political Science Data: An Alternative Algorithm for Multiple Imputation. American Political Science Review. Vol. 95, 1, 49--69 (2001)
13. Oudshoorn, C.G.M., Buuren, V.S., Rijckevorsel, V.: Flexible Multiple Imputation by Chained Equations of the AVO-95 Survey. In: TNO Prevention and Health. Report PG/VGZ/99.045 (1999)
14. Asuncion, A., Newman, D.J.: UCI Machine Learning Repository. University of California, Irvine, http://archive.ics.uci.edu/ml/datasets/Adult
15. Kohavi, R.: Scaling Up the Accuracy of Naive-Bayes Classifiers: A Decision-Tree Hybrid. Proc. In: 2-nd Int. KDDM Conf., pp. 202--207. AAAI Press, Portland (1996)
16. Ho, T.K.: The Random Subspace Method for Constructing Decision Forests. IEEE Transactions on Pattern Analysis and Machine Intelligence. Vol. 20, 8, 832--844 (1998)
17. Haykin, S.: Neural Networks: A Comprehensive Foundation, 2-nd ed. Prentice Hall (1998)
18. Gwet, K.: Statistical Tables for Inter-Rater Agreement. StatAxis, Gaithersburg (2001)
19. Van Rijsbergen, C.J.: Information Retrieval. 2-nd edition. Butterworth, London (1979)
20. Crime and Justice Research Center, Temple University, http://www.temple.edu/prodes/