Восстановление пропущенных значений в разнородных шкалах с большим числом пропусков*

О. Ю. Бахтеев
bakhteev@phystech.edu
Московский физико-технический институт

Рассматривается задача восстановления пропущенных значений в выборках, содержащих значительное число пропусков. Вводится понятие устойчивости восстановления пропуска, а также исследуется возможность применимости подхода для восстановления пропущенных значений. Исследуется случай, когда восстановление производится по k ближайшим соседям. Рассматриваются теоретические аспекты применимости данного подхода для сильно разреженных данных. Рассматривается вариант восстановления пропущенных значений с использованием восстановленных значений в качестве источника для восстановления других элементов.

Ключевые слова: *восстановление пропущенных значений; k ближайших соседей; разнородные шкалы*

Handling missing values in mixed-scale datasets with large amount of missing values*

O. Y. Bakhteev
Moscow Institute of Physics and Technology

Background: The paper investigates the problem of missing values handling in datasets with a large amount of missing values. One of the problems of missing values filling methods is their instability. The order of missing values filling can seriously change the efficiency of the method. The paper considers the case when the dataset has significant amount of features with discrete scales with low cardinality.
Methods: There are different methods of missing values handling. The paper focuses on the filling missing values using the metric properties of the dataset. The paper proposes some definitions and statements in order to formalize the problem of instability. The method using k nearest neighbors is considered. The paper considers a variation of the method that uses already filled missing values as values of nearest neighbors for new fill. Also, some theoretical aspects of this method implementation are considered.
Results: In order to analyze the behavior and efficiency of the considered method, two experiments were conducted. The results were compared with other missing values filling techniques such as filling with decision trees and filling with average value of the scale.
Concluding Remarks: The proposed mathematical framework can be used for further research of missing values filling methods.
Keywords: *imputation; missing values; k nearest neighbours; mixed-scale datasets*

1 Введение

В работе исследуется проблема восстановления значительного количества пропусков в выборке в задачах анализа данных. Основной трудностью, связанной с восстановлением

*Работа выполнена при поддержке РФФИ, грант №14-07-31045.

пропусков, является неустойчивость полученной модели при последовательном восстановлении части пропусков: порядок восстановления пропусков может значительно изменить вид восстановленной выборки. Примером данных с подобным количеством пропущенных значений является выборка историй болезни лошадей, обследованных в ветеринарной клинике [1]. В выборке присутствуют 28 признаков в разнородных шкалах, около 30% выборки заполнено пропущенными значениями. В данной работе алгоритмы восстановления пропусков исследуются на примере социологических данных [2]. В данной выборке содержится 1000 объектов с признаковым описанием в номинальных, линейных и порядковых шкалах.

Ранее был предложен ряд подходов, используемых для обработки пропущенных значений. В работе [3] рассматривается исключение из выборки данных с пропущенными значениями. При значительном количестве пропущенных значений данный метод не позволяет построить адекватную модель выборки. Кроме того в случае если в выборке не существует объекта с полностью восстановленными атрибутами, метод неприменим. В работах [3, 4] рассматривается построение математических моделей на подмножествах атрибутов, соответствующих восстановленным атрибутам объектов. Для каждого объекта выборки находится подмножество его восстановленных атрибутов, и для нее строится математическая модель. Данный подход требует согласования математических моделей, учитывающих разный набор атрибутов каждого объекта, и потому требует больших вычислительных ресурсов. Оба этих метода отбрасывают пропущенные значения в данных. Другим подходом к их обработке является восстановление пропусков [5] по имеющимся данным выборки. Перечислим некоторые его варианты:

- восстановление средними значениями атрибута по всей выборке [6]. Метод является достаточно простым для реализации, однако дает грубые результаты и может ухудшить результаты работы дальнейшей классификации или регрессии [7];
- восстановление с использованием предсказательной модели (Predictive value imputation) [3]. Данный метод предполагает восстановление пропущенного значения на основе некоторой зависимости данных исходной выборки;
- восстановление с использованием распределения значений атрибута [3,8]. Метод предполагает оценку распределения значений атрибута и дальнейшее восстановление данных с использованием этого распределения. Данный подход можно встретить, например, в алгоритме построения дерева решений C4.5.

В работах [3,6] проводится обзор основных подходов к восстановлению пропущенных значений. В работе [8] описывается подход к восстановлению, основанный на методах прикладной статистики и теории вероятностей.

В работах [9, 10] рассматривается подход множественных заполнений (Multiple imputation), основанный на методе Монте Карло. При использовании этого подхода восстановление каждого пропуска происходит несколько раз, таким образом генерируются несколько полностью восстановленных выборок. Затем происходит слияние полученных выборок.

В работах [6,11–15] рассматривается подход к восстановлению пропущенных значений, основанный на методе k ближайших соседей. Данный подход восстановления пропусков восстанавливает значения как в непрерывных шкалах, так и в дискретных [6]. В работах [6, 11, 13, 14] отражены результаты экспериментов по восстановлению пропущенных значений с применением данного подхода. В работе [15] для оценки погрешности метода k ближайших соседей использовались методы математической статистики, производились

оценки среднеквадратичного отклонения реального значения атрибута от значения, полученного методом k ближайших соседей. В работе [16] рассматривается проблема восстановления пропусков по неполностью восстановленным соседям. Частично эта проблема решается в [12], где предлагается итеративная версия восстановления пропущенных значений. На первой итерации все пропуски восстанавливаются средним значением признака.

В данной работе исследуется задача восстановления пропусков в случае значительного числа признаков, выполненных в дискретных шкалах малой мощности. В работе не вводятся статистические предположения о распределении значений признаков. Подобный класс данных встречается в задачах экспертного оценивания [17, 18].

Для восстановления пропущенных значений рассматривается подход, основанный на восстановлении пропусков по k ближайшим соседям. Вводится функция устойчивости восстановления, учитывающая, насколько восстановление может улучшить дальнейшее восстановление пропусков выборки. Предлагается подход, позволяющий проводить транзитивные восстановления [16], т. е. использование объектов с пропуском в некотором поле для дальнейшего восстановления пропусков в этом же поле для других объектов. Второй вариант алгоритма не использует транзитивное восстановление. Вводится функция ошибки восстановления пропущенных значений, соответствующая сумме расстояний до реальных значений объектов, в метрике пространства объектов. Изучаются границы применимости предлагаемого алгоритма восстановления пропущенных значений.

2 Формальная постановка задачи

В данном разделе вводятся формальная постановка задачи восстановления пропущенных значений и определения, требуемые для формализации задачи восстановления пропущенных значений.

Определение 1. *Шкала $\mathbb{L}$ — алгебраическая структура* [19] *с заданным набором операций и отношений, удовлетворяющая фиксированному набору аксиом.*

Определение 2. *Номинальная шкала $\mathbb{C}$ — шкала с заданным на ней бинарным отношением равенства:*

1. $x = y \vee x \neq y$;
2. $x, y : x = y \Rightarrow y = x$;
3. $x, y, z : x = y \wedge y = z \Rightarrow x = z$,

где x, y, z — объекты, представленные в шкале $\mathbb{C}$: $x, y, z \in \mathbb{C}$.

Определение 3. *Порядковая шкала $\mathbb{O}$ — номинальная шкала с заданным на ней бинарным отношением R, для которого выполнены следующие свойства:*

1. xRx,
2. $xRy \wedge yRx \Rightarrow x = y$;
3. $xRy \wedge yRz \Rightarrow xRz$;

где $x, y, z \in \mathbb{O}$.

Определение 4. *Линейная шкала $\mathbb{W}$ — порядковая шкала с отношением полного порядка и определенными операциями сложения и вычитания.*

Задана выборка $\mathbf{X}$ — множество вектор-строк:

$$\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_m]^\top \subset \mathbb{X},$$

$$\mathbf{X} = \begin{bmatrix}\mathbf{x}_1\\ \mathbf{x}_2\\ \mathbf{x}_3\end{bmatrix} \begin{bmatrix}1 & \square\\ \square & 2\\ 4 & 3\end{bmatrix} \to \hat{\mathbf{X}}^1 = \begin{bmatrix}1 & 3\\ \square & 2\\ 4 & 3\end{bmatrix} \to \hat{\mathbf{X}}^2 = \begin{bmatrix}1 & 3\\ 4 & 2\\ 4 & 3\end{bmatrix}$$

Рис. 1 Пример восстановления пропущенных значений

лежащих в пространстве $\mathbb{X}$:

$$\mathbb{X} = (\mathbb{L}_1 \cup \{\square\}) \times \ldots \times (\mathbb{L}_n \cup \{\square\}).$$

где $\mathbb{X}$ — множество возможных значений векторов признаков объектов или пространство объектов с введенной на нем метрикой d; $\mathbb{L}_j$ — линейная, номинальная или порядковая шкала, $\square$ — символ, соответствующий пропущенному значению. В выборке находится ℓ пропущенных значений, $\ell > 0$.

Определение 5. *Пусть $\mathbf{x}_i \in \mathbf{X}$ — объект, имеющий пропуск в j-м признаке. Пусть объекты $\mathbf{x}_{q_1}, \ldots, \mathbf{x}_{q_k} \in \mathbf{X}$ — объекты с заполненными значениями j-го признака. Операцией восстановления j-го признака объекта $\mathbf{x}_i$ по объектам $\mathbf{x}_{q_1}, \ldots, \mathbf{x}_{q_k} \in \mathbf{X}$ назовем следующее отображение:*

$$\mathbf{x}_i \leftarrow \{x_{q_1 j}, \ldots, x_{q_k j}\} = [x_{i1}, \ldots, x_{ij-1}, \operatorname{average}([x_{q_1 j}, \ldots, x_{q_k j}]), x_{ij+q}, \ldots, x_{in}],$$

где

$$\operatorname{average}([x_{q_1 j}, \ldots, x_{q_k j}]) = \begin{cases} \operatorname{mean}([x_{q_1 j}, \ldots, x_{q_k j}]), & \text{если шкала } \mathbb{L}_j \text{ — линейная;}\\ \operatorname{median}([x_{q_1 j}, \ldots, x_{q_k j}]), & \text{если шкала } \mathbb{L}_j \text{ — порядковая;}\\ \operatorname{mode}([x_{q_1 j}, \ldots, x_{q_k j}]), & \text{если шкала } \mathbb{L}_j \text{ — нормальная;}\end{cases}$$

k — множество соседей, т. е. объектов по которым восстанавливается признак.

В дальнейшем операцию восстановления $\mathbf{x}_i \leftarrow \{\mathbf{x}_{q_1 j}, \ldots, \mathbf{x}_{q_k j}\}$ будем отождествлять с кортежем вида $\mathbf{t} = (i, j, q_1, \ldots, q_k)$. Также будем обозначать через $\hat{\mathbf{X}}^b$ выборку, полученную из исходной $\mathbf{X}$ последовательным выполнением b операций восстановления. Будем полагать $\hat{\mathbf{X}}^0 = \mathbf{X}$.

Определение 6. *Операцию $\mathbf{t} = (i, j, q_1, \ldots, q_k)$ назовем корректной для выборки $\mathbf{X}$, если $x_{ij} = \square$ и $x_{q_r j} \neq \square$ для $r \in \{1, \ldots, k\}$.*

Определение 7. *Последовательность операций восстановления $\mathbf{T} = (\mathbf{t}_1, \ldots, \mathbf{t}_b)$ назовем корректной, если каждая операция $\mathbf{t}_p \in \{\mathbf{t}_1, \ldots, \mathbf{t}_b\}$ корректна для выборки $\hat{\mathbf{X}}^{p-1}$, где $\hat{\mathbf{X}}^{p-1}$ — выборка, полученная из $\mathbf{X}$ последовательным выполнением операций $\mathbf{t}_1, \ldots, \mathbf{t}_{p-1}$, $\hat{\mathbf{X}}^0 = \mathbf{X}$, $p \in \{1, \ldots, b\}$.*

Пример восстановления значений с помощью последовательности из двух операций $\mathbf{T} = ((1, 2, 3)(2, 1, 3))$ приведен на рис. 1.

Определение 8. *Множество корректных последовательностей операций восстановления длины ℓ обозначим как $\mathbf{C}_\ell$.*

Определение 9. *Обозначим за $\operatorname{filled}_o(\mathbf{x}, \mathbf{X})$ множество индексов заполненных значений объекта $\mathbf{x}$ в выборке $\mathbf{X}$. Обозначим за $\operatorname{filled}_f(j, \mathbf{X})$ множество индексов объектов с заполненным признаком j в выборке $\mathbf{X}$.*

Определение 10. *Пусть* $\mathbf{T} = (\mathbf{t}_1, \ldots, \mathbf{t}_b)$ *— корректная последовательность операций,* $\mathbf{t}_b = \{i, j, q_1, \ldots, q_k\}$, $|\mathbf{T}| \geqslant 0$. *Устойчивостью восстановления* $x_{ij} \leftarrow \mathbf{x}_{q_1}, \ldots, \mathbf{x}_{q_k}$ *под действием* $\mathbf{T}$ *назовем следующую величину:*

$$u(\mathbf{x}_i \leftarrow x_{q_1 j}, \ldots, x_{q_k j}|\mathbf{T}) = \mathrm{mean}_{r\in\{1,\ldots,k\}} \left(\frac{|\mathrm{filled}_o(\hat{\mathbf{x}}_i^b, \mathbf{X}^b) \cap \mathrm{filled}_o(\hat{\mathbf{x}}_{q_r}^b, \mathbf{X}^b)|}{n} \frac{|\mathrm{filled}_f(j, \mathbf{X}^b)}{m} \right).$$

Определение 11. *Пусть* $\mathbf{T} = (\mathbf{t}_1, \ldots, \mathbf{t}_l)$ *— корректная последовательность длины* ℓ, $\mathbf{t}_b = (i_b, j_b, q_{(b,1)}, \ldots, q_{(b,k)})$, *где индекс* b *пробегает от* 1 *до* ℓ, $b \in \{1, \ldots, \ell\}$.

Устойчивостью последовательности восстановлений $\mathbf{T}$ *для выборки* $\mathbf{X}$ *назовем величину:*

$$U(\mathbf{X}|\mathbf{T}) = u(\mathbf{x}_{i_1} \leftarrow x_{q_{(1,1)}j_1}, \ldots, x_{q_{(1,k)}j_1}) + \sum_{b=2}^{\ell} u(\mathbf{x}_{i_b} \leftarrow x_{q_{(b,1)}j_b}, \ldots, x_{q_{(b,k)}j_b}|(\mathbf{t}_1, \ldots, \mathbf{t}_{b-1})).$$

Требуется найти последовательность восстановлений $\mathbf{T}$, решающую следующую задачу оптимизации:

$$\mathbf{T} = \underset{\mathbf{T}'\in\mathbf{C}_\ell, U(\mathbf{X}|\mathbf{T}')=\max}{\arg\min} \sum_{i=1}^{m}\sum_{j=1}^{n} d(\hat{x}_{ij}, x'_{ij}),$$

где $\hat{\mathbf{x}}_i = [x_{i1}, \ldots, x_{in}]$ — объект, восстановленный под действием последовательности $\mathbf{T}$; $\mathbf{x}'$ — объект с реальными значениями пропусков; d — метрика на пространстве $\mathbb{X}$. Подробно метрики в разнородных шкалах описаны в разд. 5.

Таким образом, исходная задача разбивается на две подзадачи:

1. Нахождение множества корректных последовательностей $\mathbf{T}'$, под действием которых устойчивость U выборки максимальна.
2. Выбор последовательности $\mathbf{T}$, доставляющей минимум сумме расстояний от восстановленных объектов до реальных.

Рассмотрим подробнее операцию восстановления пропущенных значений. Будем восстанавливать пропущенные значения с восстановлением пропусков по k ближайшим соседям. При использовании данного алгоритма пропущенные значения объекта выборки восстанавливаются по множеству k ближайших к нему объектов выборки. Рассмотрим данный алгоритм восстановления пропусков на примере.

2.1 Пример 1

Пусть задана выборка $\mathbf{X}$:

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \mathbf{x}_3 \end{bmatrix} \begin{bmatrix} 1 & 2 & \square & \square & \square \\ 1 & 2 & 6 & 5 & \square \\ 4 & 8 & 0 & \square & 5 \end{bmatrix}.$$

Будем считать, что на пространстве $\mathbb{X}$ и на каждом его подпространстве $\mathbb{X}_\mathcal{J} = \prod_{j\in\mathcal{J}} \mathbb{L}_j$ задана метрика $d_\mathcal{J}$, принимающая значения из отрезка $[0; 1]$.

Восстановим пропущенное значение x_{13} с применением алгоритма k ближайших соседей. Рассмотрим случай $k = 1$.

Для объектов $\mathbf{x}_1$, $\mathbf{x}_2$ и $\mathbf{x}_3$ имеется общее подпространство $\mathbb{X}_\mathcal{J}$, т. е. такое пространство, в котором ни у одного объекта не содержится пропущенных значений. Это подпространство соответствует проекции пространства объектов $\mathbb{X}$ на первые два признака. Определим, какой из объектов $\mathbf{x}_2$, $\mathbf{x}_3$ является ближайшим для объекта $\mathbf{x}_1$:

$$\mathbf{x}' = \underset{\mathbf{x}'\in\{\mathbf{x}_2,\mathbf{x}_3\}}{\arg\min}\ d_\mathcal{J}\big(\mathrm{f}_\mathcal{J}(\mathbf{x}_1), \mathrm{pr}_\mathcal{J}(\mathbf{x}')\big), \tag{1}$$

где $d_{\mathcal{J}}$ — метрика на пространстве $\mathbb{X}_{\mathcal{J}}$, принимающая значения из $[0;1]$; $\mathrm{pr}_{\mathcal{J}}$ — функция, проецирующая объекты на пространство $\mathbb{X}_{\mathcal{J}}$. Пусть согласно метрике $d_{\mathcal{J}}$ $\mathbf{x}' = \mathbf{x}_2$.

Восстановим пропущенный признак x_{13} по ближайшему соседу $\mathbf{x}'$:

$$\hat{\mathbf{x}}_1 = \mathbf{x}_1 \leftarrow \{x_{23}\}.$$

В случае $k > 1$ восстановленное значение усредняется по нескольким ближайшим соседям. Так, в данном примере при $k = 2$ восстановленное значение x_{13} будет равняться среднему соответствующих значений объектов $\mathbf{x}_2, \mathbf{x}_3$.

Рассмотрим теперь случай, когда для всех трех объектов $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3$ не существует общего пространства $\mathbb{X}_{\mathcal{J}}$.

2.2 Пример 2

Пусть задана выборка $\mathbf{X}$:

$$\mathbf{X} = \begin{bmatrix} 1 & 2 & \square & \square & \square \\ \square & 2 & 6 & 5 & \square \\ 4 & \square & 0 & \square & 5 \end{bmatrix}.$$

Здесь общим пространством для объектов $\mathbf{x}_1, \mathbf{x}_2$ является подпространство $\mathbb{X}_{\{2\}}$, содержащее только второй признак, для объектов $\mathbf{x}_1, \mathbf{x}_3$ — подпространство $\mathbb{X}_{\{1\}}$, содержащее только первый признак.

В данном случае ближайший сосед будет определяться по различным подпространствам. Если метрики принимают значения из одного множества, будем находить ближайших соседей в различных пространствах, сравнивая полученные значения метрик между собой.

Предлагается использовать для восстановления пропущенных значений транзитивное восстановление, т. е. восстановление с использованием объектов с незаполненными полями в качестве соседей для восстановления этого же поля для других объектов. Поясним данный момент на следующем примере.

2.3 Пример 3

Пусть задана выборка $\mathbf{X}$:

$$\mathbf{X} = \begin{bmatrix} 1 & \square & \square \\ 1 & \square & 3 \\ \square & 2 & 3 \end{bmatrix}.$$

В данном примере для восстановления пропуска x_{12} требуется сперва восстановить пропуск x_{22} по соседу $\mathbf{x}_3$. Разрешение такого транзитивного восстановления усложняет алгоритм, однако при этом позволяет восстановить более широкий класс данных.

3 Нахождение оптимальной последовательности восстановления пропусков

Для построения стратегии восстановления пропусков будем использовать аппроксимацию функции устойчивости, которая не будет зависеть от соседей, по которым восстанавливается объект. Для дальнейшего рассмотрения задачи введем ряд определений.

Определение 12. *Пусть $\mathbf{x}_i \in \mathbf{X}$ — объект, имеющий пропуск в j-м признаке. Абстрактной операцией восстановления j-го признака объекта $\mathbf{x}_i$ назовем множество всех возможных операцией восстановления данного признака по одному соседу:*

$$\mathbf{x}_i \leftharpoondown j = \{(\mathbf{x}_i \leftarrow \{x_{qj}\}), q \in \mathrm{filled}_f(j, \mathbf{X})\}.$$

По аналогии с операцией восстановления абстрактную операцию восстановления $\mathbf{x}_i \leftharpoondown \leftharpoondown j$ будем отождествлять с кортежем (i, j).

Определение 13. *Последовательность абстрактных операций восстановления* $\bar{\mathbf{T}} = (\bar{\mathbf{t}}_1, \bar{\mathbf{t}}_b), \bar{\mathbf{t}} = (i, j)$, *где* $i \in \{1, \ldots, m\}, j \in \{1, \ldots, n\}$, *будем называть корректной последовательностью абстрактных операций восстановления для выборки* $\mathbf{X}$, *если для каждого* $\bar{\mathbf{t}} = (i, j)$: $x_{ij} = \square$ *и каждая пара кортежей* $\bar{\mathbf{t}}_1, \bar{\mathbf{t}}_2 \in \bar{\mathbf{T}}$ *отличается хотя бы по одной координате.*

Определение 14. *Пусть* $\bar{\mathbf{T}} = (\bar{\mathbf{t}}_1, \ldots, \bar{\mathbf{t}}_b)$ — *корректная последовательность абстрактных операций восстановления,* $|\bar{\mathbf{T}}| \geqslant 0$. *Пусть* $\bar{\mathbf{X}}^b$ — *выборка, полученная из исходной выборки* $\mathbf{X}$ *восстановлением значений* $x_{ij}, (i, j) \in \bar{\mathbf{T}}$ *произвольными значениями соответствующей шкалы,* $\bar{\mathbf{x}}_i^b \in \bar{\mathbf{X}}^b$. *Аппроксимированной устойчивостью пропуска* x_{ij} *назовем следующую величину:*

$$\bar{u}(x_{ij}|\bar{\mathbf{T}}) = \frac{|\text{filled}_o(\bar{\mathbf{x}}_i^b, \bar{\mathbf{X}}^b)|}{n} \frac{|\text{filled}_f(j, \bar{\mathbf{X}}^b)|}{m}.$$

Данная функция, в отличие от функции устойчивости u, не учитывает пересечение множества заполненных признаков восстанавливаемого объекта $\mathbf{x}_i$ и объектов, по чьим значениям восстанавливается пропуск. Таким образом, аппроксимированная устойчивость является верхней оценкой функции u.

Теорема 1. *Для каждого* $i \in \{1, \ldots, m\}, j \in \{1, \ldots, n\}$, *любого множества объектов* $\mathbf{x}_{q_1}, \ldots, \mathbf{x}_{q_k}$, *имеющих заполненный признак* j *и корректной последовательности операций восстановления* $\mathbf{T}, |\mathbf{T}| = |\bar{\mathbf{T}}|$, *такой, что для любой операции* $\mathbf{t}_p \in \mathbf{T}$ *первые два элемента кортежа* $\mathbf{t}_p$ *равны элементам кортежа* $\bar{\mathbf{t}}_p \in \bar{\mathbf{T}}$, *следует, что*

$$u(\mathbf{x}_i \leftharpoondown \mathbf{x}_{q_1 j}, \ldots, \mathbf{x}_{q_k j}|\mathbf{T}) \leqslant \bar{u}(x_{ij}|\bar{\mathbf{T}}).$$

Доказательство. Доказательство следует из определений устойчивости заполнения и аппроксимированной устойчивости:

$$u(\mathbf{x}_i \leftharpoondown \mathbf{x}_{q_1 j}, \ldots, \mathbf{x}_{q_k j}|\mathbf{T}) = \text{mean}_{r \in \{1,\ldots,k\}} \frac{|\text{filled}_o(\hat{\mathbf{x}}_i^b, \hat{\mathbf{X}}^b) \cap \text{filled}_o(\hat{\mathbf{x}}_{q_r}^b, \hat{\mathbf{X}}^b)|}{n} \frac{|\text{filled}_f(j, \hat{\mathbf{X}}^b)|}{m} \leqslant$$
$$\leqslant \frac{|\text{filled}_o(\bar{\mathbf{x}}_i^b, \bar{\mathbf{X}}^b)|}{n} \frac{|\text{filled}_f(j, \bar{\mathbf{X}}^b)|}{m} = \bar{u}(x_{ij}|\bar{\mathbf{T}}).$$

■

По аналогии с определением устойчивости последовательности операций восстановления системы введем понятие аппроксимированной устойчивости последовательности абстрактных операций восстановления.

Определение 15. *Пусть* $\bar{\mathbf{T}} = (\bar{\mathbf{t}}_1, \ldots, \bar{\mathbf{t}}_b)$ — *последовательностей корректных абстрактных операций восстановления.* b*-Аппроксимированной устойчивостью последовательности абстрактных операций восстановления для выборки* $\mathbf{X}$ *назовем следующую величину:*

$$\bar{U}^b(\mathbf{X}|\bar{\mathbf{T}}) = \bar{u}(x_{i_1 j_1}) + \sum_{r=1}^{b-1} \bar{u}(x_{i_{r+1} j_{r+1}}|\bar{\mathbf{t}}_1, \ldots, \bar{\mathbf{t}}_r) + \sum_{(i,j) \notin \bar{\mathbf{T}}, x_{ij} = \square} \bar{u}(x_{ij}|\bar{\mathbf{t}}_1, \ldots, \bar{\mathbf{t}}_b).$$

Из определения и предыдущей теоремы немедленно вытекает следующее утверждение.

Теорема 2. *Пусть* $\mathbf{T} = (\mathbf{t}_1, \ldots, \mathbf{t}_\ell)$ *— корректная последовательность длины* ℓ; $\bar{\mathbf{T}} = (\bar{\mathbf{t}}_1, \ldots, \bar{\mathbf{t}}_\ell)$ *— корректная последовательность абстрактных операций восстановления длины* ℓ, *такая что для любой операции* $\mathbf{t}_p \in \mathbf{T}$ *первые два элемента кортежа* $\mathbf{t}_p$ *равны элементам кортежа* $\bar{\mathbf{t}}_p \in \bar{\mathbf{T}}$. *Тогда*

$$U(\mathbf{X}|\mathbf{T}) \leqslant \bar{U}^\ell(\mathbf{X}|\bar{\mathbf{T}}).$$

Вместо исходной задачи максимизации устойчивости системы $U(\mathbf{X}|\mathbf{X})$ будем оптимизировать аппроксимацию $\bar{U}(\mathbf{X}|\bar{\mathbf{T}})$. На каждом шаге итерации алгоритм должен отбирать пропуск x_{ij}, имеющий корректную операцию восстановления $\mathbf{t}$, дающую максимум устойчивости $u(x_{ij} \leftarrow (\mathbf{x}_{q_1}, \ldots, \mathbf{x}_{q_k}))$. Для учета последующих шагов восстановления будем просматривать аппроксимированную устойчивость выборки на несколько шагов вперед.

Для вычисления аппроксимированной b-устойчивости определим понятие графа зависимостей, соответствующего выборке $\mathbf{X}$.

Определение 16. *Графом зависимости* $\langle \mathbf{V}, \mathbf{E} \rangle$, *соответствующим выборке* $\mathbf{X}$, *назовем совокупность вершин и ребер, где каждая вершина* v_{ij} *соответствует элементу* x_{ij}, *а ребра строятся по следующим правилам:*

- *если в объекте* $\mathbf{x}_i$ *существует два пропущенных значения* x_{ij_1}, x_{ij_2}, *то между вершинами* v_{ij_1}, v_{ij_2} *существует ребро* e_{ij_1,ij_2};
- *если в объектах* $\mathbf{x}_{i_1}, \mathbf{x}_{i_2}$ *пропущено значение* j*-го признака, то между вершинами* v_{i_1}, v_{i_2} *существует ребро* $e_{i_1 j, i_2 j}$.

Приведем пример графа зависимости.

3.1 Пример 4

Пусть задана выборка

$$\mathbf{X} = \begin{bmatrix} 1 & \square & \square \\ 1 & \square & 3 \\ 1 & \square & 3 \\ \square & 2 & 3 \end{bmatrix}.$$

Граф зависимости для данной выборки изображен на рис. 2.

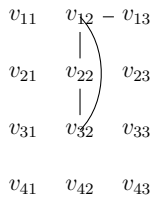

Рис. 2 Граф, соответствующий выборке $\mathbf{X}$

Определение 17. *Объектной степенью вершины* $\deg_O(v)$ *назовем степень вершины* v *с учетом только ребер между вершинами, соответствующими одному объекту.*
Признаковой степенью вершины $\deg_F(v)$ *назовем степень вершины* v *с учетом только ребер между вершинами, соответствующими одному признаку.*

Докажем ряд утверждений для реализации жадной стратегии выбора операции восстановления.

Теорема 3. *Пусть* $\bar{\mathbf{T}} = (\bar{\mathbf{t}}_1, \ldots, \bar{\mathbf{t}}_b)$ *— последовательность абстрактных операций восстановления,* $\bar{\mathbf{t}}_r = (i_r, j_r), r \in \{1, \ldots, b\}$. *Аппроксимированная устойчивость пропуска* x_{ij} *при условии последовательности* $\bar{\mathbf{T}}$ *выглядит следующим образом:*

$$\bar{u}(x_{ij}|\mathbf{t}_1, \ldots, \mathbf{t}_b) = \bar{u}(x_{ij}) + \delta(x_{ij}, x_{i_1 j_1}|\varnothing) + \sum_{r=2,\ldots,b} \delta(x_{ij}, x_{i_r j_r}|\bar{\mathbf{t}}_1, \ldots, \bar{\mathbf{t}}_{r-1}).$$

Здесь

$$\delta(x_{ij}, x_{qr}|\mathbf{I}) = \begin{cases} 0, \text{ если } i \neq q \text{ и } j \neq k; \\ \dfrac{(n - \deg_O(v_{ij}) - 1 + |O(i, \mathbf{I})|)}{mn}, \text{ если } j = r; \\ \dfrac{(m - \deg_F(v_{ij}) - 1 + |F(j, \mathbf{I})|)}{mn}, \text{ если } i = q, \end{cases}$$

где $O(i, \mathbf{I})$ *— множество кортежей из* $\mathbf{I}$*, на первом месте которых стоит* i*;* $F(j, \mathbf{I})$ *— множество кортежей из* $\mathbf{I}$*, на втором месте которых стоит* j*.*

Доказательство.

Пусть для начала $b = 0$. Тогда равенство является тривиальным.

Пусть теперь $b = 1$. Рассмотрим изменение аппроксимированной устойчивости $\Delta = \bar{u}(x_{ij}|\mathbf{t}_1) - \bar{u}(x_{ij})$ при условии кортежа $\bar{\mathbf{t}}_1$. Если $i \neq i_1$ и $j \neq j_1$, то восстановление пропуска $x_{i_1 j_1}$ никак не влияет на аппроксимированную устойчивость x_{ij}. Если $i = i_1$, то filled_o увеличится на единицу и, следовательно, аппроксимированная устойчивость увеличится на $|\text{filled}_f(j, \mathbf{X})|/(mn) = (m - \deg_F(v_{ij}) - 1)/(mn)$ относительно величины $\bar{u}(x_{ij})$. Если $j = j_1$, то filled_f увеличится на единицу и, следовательно, аппроксимированная устойчивость увеличится на $|\text{filled}_o(\mathbf{x}_i, \mathbf{X})|/(mn) = (n - \deg_O(v_{ij}) - 1)/(mn)$ относительно величины $\bar{u}(x_{ij})$, и равенство выполняется.

В случае $b > 1$ доказательство производится аналогично. Рассмотрим изменение аппроксимированной устойчивости для каждого $r \in \{1, \ldots, b\}$: $\Delta = \bar{u}(x_{ij}|\mathbf{t}_1, \ldots, \mathbf{t}_r) - \bar{u}(x_{ij}|\mathbf{t}_1, \ldots, \mathbf{t}_{r-1})$ при условии кортежей $\bar{\mathbf{t}}_1, \ldots, \bar{\mathbf{t}}_r$. Аналогично случаю $b = 1$ Δ может измениться на $|\text{filled}_f(j, \bar{\mathbf{X}}^{r-1})|/(mn) = \delta(x_{ij}, x_{i_r j_r}|\bar{\mathbf{t}}_1, \ldots, \bar{\mathbf{t}}_{r-1})$ или $|\text{filled}_o(\mathbf{x}_i, \bar{\mathbf{X}}^{r-1})|/(mn) = \delta(x_{ij}, x_{i_r j_r}|\bar{\mathbf{t}}_1, \ldots, \bar{\mathbf{t}}_{r-1})$, и равенство выполняется. ■

Рассмотрим частный случай 1-аппроксимированной устойчивости выборки.

Теорема 4. *Пусть задан кортеж* $\bar{t} = (i, j)$*, такой что* $i \in \{1, \ldots, m\}, j \in \{1, \ldots, n\}$*,* $x_{ij} = \square$*. Тогда аппроксимированная* 1*-устойчивость выборки при условии* t *будет равняться:*

$$\bar{U}^1(\mathbf{X}|\mathbf{t}) = \bar{U}^0(\mathbf{X}) + \frac{\deg_O(v_{ij})}{n} + \frac{\deg_F(v_{ij})}{m} - \sum_{e_{ij,ij_2} \in \mathbf{E}} \frac{\deg_F(v_{ij_2}) + 1}{mn} - \sum_{e_{ij,i_2 j} \in \mathbf{E}} \frac{\deg_O(v_{i_2 j}) + 1}{mn}.$$

Доказательство. Всего существует $\deg_O(v_{ij})$ пропусков в объекте $\mathbf{x}_i$ и $\deg_F(v_{ij})$ пропусков в объектах в признаке j, не считая пропуск u_{ij}. Суммируя аппроксимированную устойчивость всех пропусков и группируя значения функции δ по пропускам в объекте i и в признаке j, получаем требуемое равенство. ■

4 Формализация рассматриваемого алгоритма

Формализуем полученный алгоритм. Для дальнейшего описания алгоритма введем понятие разрешимого пропуска, т. е. пропуска, который может быть восстановлен алгоритмом. В терминах предложенного алгоритма данное понятие определяется следующим образом.

Вход: Выборка $\mathbf{X}$ с пропущенными значениями; число соседей k; длина аппроксимации b;
Выход: Восстановленная выборка $\hat{\mathbf{X}}$;

1: **пока** множество разрешимых пропусков $\mathbf{R}$ не пусто
2: $x_{ij} = \underset{x_{i_1 j_1}, \bar{\mathbf{t}}_2, \ldots, \bar{\mathbf{t}}_b}{\arg\max} \bar{U}^b(\mathbf{X}|(i_1, j_1), \bar{\mathbf{t}}_2, \ldots, \bar{\mathbf{t}}_b))$;
3: Для пропуска x_{ij} получить соседей $\mathbf{N}$;
4: Упорядочить $\mathbf{N}$ в лексикографическом порядке по устойчивости восстановления x_{ij} и расстоянию до $\mathbf{x}_i$;
5: Получить первые k объектов $\mathbf{x}_{q_1}, \ldots, \mathbf{x}_{q_k}$ из упорядоченного множества соседей;
6: $\hat{\mathbf{x}}_i = \mathbf{x}_i \leftarrow (x_{q_1 j}, \ldots, x_{q_k j})$;

Рис. 3 Псевдокод предложенного алгоритма восстановления

Определение 18. *Пропуск x_{ij} является разрешимым, если существуют объекты $\mathbf{x}_{q_1}, \ldots, \mathbf{x}_{q_k}$, каждый из которых имеет непустой признак j, а также хотя бы один общий заполненный признак с $\mathbf{x}_i$, т. е.* $\text{filled}_O(\mathbf{x}_i, \mathbf{X}) \cap \text{filled}_O(\mathbf{x}_{q_r}, \mathbf{X}) \neq \varnothing, r \in \{1, \ldots, k\}$.

Из определения разрешимого пропуска следует, что для каждого разрешимого пропуска x_{ij} существуют объекты $\mathbf{x}_{q_1}, \ldots, \mathbf{x}_{q_k}$ такие, что устойчивость операции восстановления данного x_{ij} по этим объектам больше нуля:

$$u(x_{ij} \leftarrow (\mathbf{x}_{q_1}, \ldots, \mathbf{x}_{q_k})) > 0\,.$$

Пусть задано число соседей k и длина аппроксимации b. На каждой итерации алгоритма будем отбирать множество разрешимых пропусков $\mathbf{R}$.

Из множества разрешимых пропусков выберем такой пропуск x_{ij}, для которого b-аппроксимированная устойчивость пропуска максимальна, т. е. $x_{ij} = \underset{x_{i_1 j_1}, \bar{\mathbf{t}}_2, \ldots, \bar{\mathbf{t}}_b}{\arg\max} \bar{U}^b(\mathbf{X}|(i_1, j_1), \bar{\mathbf{t}}_2, \ldots, \bar{\mathbf{t}}_b))$.

Для полученного пропуска x_{ij} получим всех соседей, т. е. такие объекты $\mathbf{N}$, что $\mathbf{x}_{qj} \neq \square$ и $\text{filled}_O(\mathbf{x}_q) \cap \text{filled}_O(\mathbf{x}_q) \neq \varnothing$, где $\mathbf{x}_q \in \mathbf{N}$.

Упорядочим объекты из $\mathbf{N}$ в лексиографическом порядке по устойчивости восстановления x_{ij} и расстоянию до $\mathbf{x}_i$:

$$\mathbf{x}_{q_1} \prec \mathbf{x}_{q_2}, \text{ если } \left[\begin{array}{c} u(x_{ij} \leftarrow \mathbf{x}_{q_1}) > u(x_{ij} \leftarrow \mathbf{x}_{q_2}); \\ u(x_{ij} \leftarrow \mathbf{x}_{q_1}) = u(x_{ij} \leftarrow \mathbf{x}_{q_2}) \text{ и } d(\mathbf{x}_i, \mathbf{x}_{q_1}) < d(\mathbf{x}_i, \mathbf{x}_{q_2}). \end{array}\right.$$

Восстановим пропущенное значение x_{ij} по k первым объектам полученного упорядоченного множества.

Псевдокод представленного алгоритма показан на рис. 3. Сложность алгоритма оценивается как $O(\ell(\ell^{b+1} + km^2))$, что намного больше сложности алгоритма без транзитивного восстановления пропусков.

5 Функции расстояния для разнородных шкал

В данном разделе проводится краткий обзор функций расстояния для различных типов шкал — линейной, порядковой, а также смешанной. Предлагается функция расстояния для выборок, описанных в линейных, номинальных и порядковых шкалах с заданным на них полным порядком.

5.1 Функция расстояния для линейных шкал

Рассмотрим обобщенную функцию расстояния для множества объектов с введенной линейной шкалой:

$$r(\mathbf{x}_i, \mathbf{x}_q) = \left(\left(|\mathbf{x}_i - \mathbf{x}_q|^p \right)^{\mathrm{T}} \mathbf{S}^{-1} |\mathbf{x}_i - \mathbf{x}_q|^p \right)^{1/(2p)},$$

где p — некоторое число; $\mathbf{S}$ — симметричная неотрицательно определенная матрица, например единичная матрица $\mathbf{I}$, а возведение вектора в степень понимается как покомпонентное возведение, т. е. $\mathbf{x}^p = (x_1^p, \ldots, x_n^p)$.

В табл. 1 представлены соответствия представленной функции различным именным функциям расстояния при фиксированных параметрах.

Таблица 1 Соответствие функции расстояния именным функциям расстояния

p	$\mathbf{S}$	Название функции	Формула
1	—	Расстояние Махаланобиса	$r(\mathbf{x}_i, \mathbf{x}_j) = \sqrt{(\mathbf{x}_i - \mathbf{x}_q)^{\mathrm{T}} \mathbf{S}^{-1} (\mathbf{x}_i - \mathbf{x}_q)}$
—	$\mathbf{I}$	Расстояние Минковского	$r(\mathbf{x}_i, \mathbf{x}_q) = \left(\sum_{k=j}^{n} |x_{ij} - x_{qj}|^q \right)^{1/q}, q = 2p$
1	$\mathbf{I}$	Евклидова Метрика	$r(\mathbf{x}_i, \mathbf{x}_q) = \sqrt{\sum_{j=1}^{n} (x_{ij} - x_{qj})^2}$
0,5	$\mathbf{I}$	Расстояние городских кварталов	$r(\mathbf{x}_i, \mathbf{x}_q) = \sum_{j=1}^{n} |x_{ij} - x_{qj}|$
$+\infty$	$\mathbf{I}$	Расстояние Чебышёва	$r(\mathbf{x}_i, \mathbf{x}_q) = \max\limits_{j=1\ldots n} (x_{ij} - x_{qj})$

5.2 Функция расстояния для порядковых шкал

Введем матричные функции $\mathbf{H}^{q+}$ и $\mathbf{H}^{q-}$ для проекции множества объектов $\mathbf{X}$ на q-й признак, где соответствующая шкала $\mathbb{L}_q$ — порядковая. Каждая компонента вектора $\mathbf{H}_i^{q+}$ определяет отношение порядка q-го признака i-го объекта с остальными объектами выборки:

$$\left(\mathbf{H}_i{}^{j+}\right)_q = \begin{cases} 1, & \text{если } x_{ij} \succ x_{qj}; \\ 0 & \text{иначе;} \end{cases}$$

$$\left(\mathbf{H}_i{}^{j-}\right)_q = \begin{cases} 1, & \text{если } x_{qj} \succ x_{ij}; \\ 0 & \text{иначе.} \end{cases}$$

Так как $||\mathbf{H}_j^{q+}||_2^2 + ||\mathbf{H}_j^{q-}||_2^2 \leqslant m$, введем функцию расстояния pdist:

$$\mathrm{pdist}(x_{iq}, x_{qj}) = \frac{m - \left(\langle \mathbf{H}_i{}^{j+}, \mathbf{H}_q{}^{j+} \rangle + \langle \mathbf{H}_i{}^{j-}, \mathbf{H}_q{}^{j-} \rangle \right)}{m}, \tag{2}$$

где m — множество объектов в выборке. Функция принимает значения из диапазона $[0;1]$.

5.3 Обобщение функции расстояния HEOM

Дополним функцию HEOM [20] для случая объектов, описанных как в номинальных и линейных шкалах, так и в порядковых шкалах с полным порядком:

$$d(\mathbf{x}_i, \mathbf{x}_q) = \frac{1}{\sqrt{n}} \left(\sum_{j=1}^{n} r(x_{ij}, x_{qj})^2 \right)^{1/2}. \tag{3}$$

Здесь

$$r(x_{ij}, x_{qj}) = \begin{cases} \text{overlap}(x_{ij}, x_{qj}), & \text{если } \mathbb{L}_j \text{ — номинальный признак;} \\ \text{pdist}(x_{ij}, x_{qj}), & \text{если } \mathbb{L}_j \text{ — порядковый признак;} \\ \text{diff}(x_{ij}, x_{qj}) & \text{иначе,} \end{cases}$$

где

$$\text{overlap}(x_{ij}, x_{qj}) = \begin{cases} 1, & \text{если } x_{ij} \neq x_{qj}; \\ 0 & \text{иначе;} \end{cases}$$

$$\text{diff}(x_{ij}, x_{qj}) = \frac{|x_{ij} - x_{qj}|}{\max\limits_{\mathbb{L}_j} - \min\limits_{\mathbb{L}_j}},$$

т. е. функция $\text{diff}(x_{ij}, x_{qj})$ определяется как нормированный модуль разницы между значениями j-го признака двух объектов.

Таким образом, мы получили метрику для смешанных шкал. Функция d принимает значения из отрезка $[0; 1]$.

6 Вычислительный эксперимент

Основной целью вычислительного эксперимента является определение границы применимости предложенного метода. С этой целью было проведено два эксперимента. В обоих экспериментах в качестве исходных данных использовалась выборка кредитозаемщиков Германии [2]. В выборке присутствует 1000 объектов и 21 признак в линейных, номинальных и порядковых шкалах. В каждом эксперименте производилась генерация подвыборки мощностью 100 объектов и добавление в нее пропущенных значений, при этом не допускалось такое добавление пропусков, при котором какой-либо объект имел бы пустое описание. Исходный код экспериментов доступен по адресу [21].

В первом эксперименте исследовалось количество неразрешимых пропусков при использовании транзитивного восстановления и без. Результаты данного эксперимента показаны на рис. 4. По оси Y отложен процент неразрешимых пропусков, по оси X — процент добавленных пропусков. Было проведено 40 запусков, результат был усреднен. Как видно из результатов, оба алгоритма могут разрешить все пропуски при достаточно большом проценте пропущенных значений.

Во втором эксперименте исследовалась эффективность рассматриваемого алгоритма восстановления пропусков. В качестве критерия ошибки Q использовалось среднее расстояние от реальных объектов до восстановленных вариантов:

$$Q = \sum_{\mathbf{x}_i \in \mathbf{X}, \exists j : \hat{\mathbf{x}}_{ij} = \square} d(\mathbf{x}_i, \hat{\mathbf{x}}_i) \cdot \frac{1}{R},$$

где $\hat{\mathbf{x}}_i$ — объект, восстановленный методом k ближайших соседей, $k = 1$; R — количество объектов, имеющих пропущенные значения.

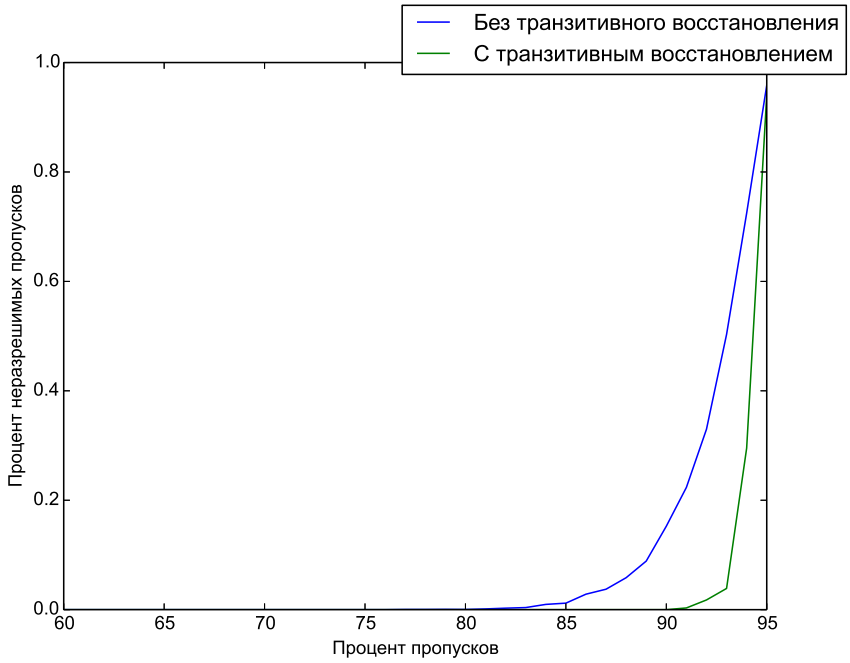

Рис. 4 Результаты первого эксперимента

Результаты данного эксперимента показаны на рис. 5. По оси X отложен процент добавленных пропусков, по оси Y — средняя ошибка восстановления. Было проведено 20 запусков, результат был усреднен. В эксперименте рассматривались алгоритм восстановления пропусков по k ближайшим соседям без транзитивного восстановления, итеративная версия алгоритма, описанная в [12], 0- и 1-аппроксимации, а также восстановление пропущенных значений средним и алгоритм восстановления с использованием дерева решений. В качестве критерия остановки итеративной версии алгоритма использовалось правило:

$$S = [\text{mean}_{\mathbf{x}\in\mathbf{X}}(d(\hat{\mathbf{x}}^u, \hat{\mathbf{x}}^{u+1}) < 0{,}01)]\,,$$

где $\hat{\mathbf{x}}^u$ — объект, восстановленный на итерации u. Как видно из результатов, наилучший результат был показан алгоритмом восстановления пропусков с использованием дерева решений. 0- и 1-аппроксимации показали результат, близкий к исходному алгоритму без транзитивного восстановления, при этом 1-аппроксимация в целом оказалась менее эффективна, чем 0-аппроксимация.

7 Заключение

В работе была рассмотрена проблема восстановления пропущенных значений в разнородных шкалах в случае значительного количества пропусков. Для формализации рассмотренной проблемы было введено понятие устойчивости восстановления пропуска и устойчивости восстановления выборки. Были рассмотрены варианты алгоритма заполнения пропусков по k ближайшим соседям, а также теоретические аспекты их применимости. Для оценки качества рассмотренных алгоритмов был проведен вычислительный эксперимент со сравнением данных алгоритмов с заполнением средними значениями и алгоритмом заполнения по дереву решений. Эксперимент показал, что наилучший результат достигается алгоритмом заполнения с использованием дерева решений.

Автор выражает благодарность д. ф.-м. н. Вадиму Викторовичу Стрижову за постановку задачи и внимание к работе.

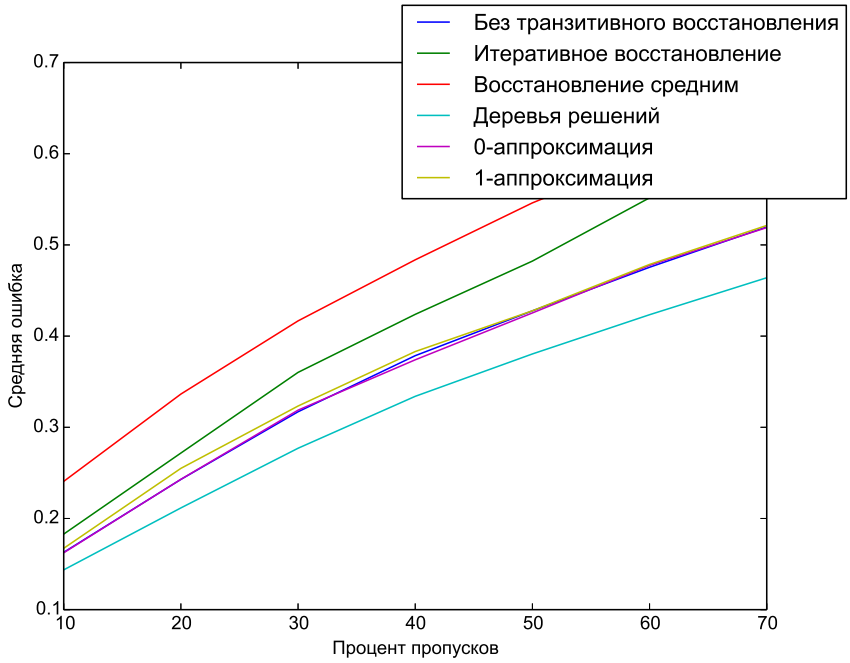

Рис. 5 Результаты второго эксперимента

Литература

[1] Horse Colic Data Set. `https://archive.ics.uci.edu/ml/datasets/Horse+Colic`.

[2] `http://archive.ics.uci.edu/ml/machine-learning-databases/statlog/german/`.

[3] *Saar-Tsechansky M., Provost F.* Handling missing values when applying classification models // J. Machine Learning Res. Arch., 2007. Vol. 8. P. 1623–1657.

[4] *Sharpe P. K., Solly R. J.* Dealing with missing values in neural network-based diagnostic systems // Neural Comput. Appl., 1995. Vol. 3. No. 2. P. 73–77.

[5] *Saunders J. A., Morrow-Howell N., Spitznagel E., Dori P., Proctor E. K., Pescarino R.* Imputing missing data: A comparison of methods for social work researchers // Social Work Res., 2006. Vol. 30. No. 1. P. 19–31.

[6] *Batista G., Monard M. C.* A study of k-nearest neighbour as an imputation method // 2nd Conference (International) on Hybrid Intelligent Systems Proceedings. Santiago, Chile: IOS Press, 2002. P. 251–260.

[7] *Durrant G. B.* Imputation methods for handling item-nonresponse in the social sciences: A methodological review, 2005. http://eprints.ncrm.ac.uk/86/1/MethodsReviewPaperNCRM-002.pdf.

[8] *Marlin B. M.* Missing data problems in machine learning. — Toronto: University of Toronto, 2008. PhD Thesis. 164 p.

[9] *Shapfer J.L.* Multiple imputation: A primer // Stat. Meth. Medical Res., 1999. Vol. 81. No. 1. P. 3–15.

[10] *Bouhlila D. S., Sellaouti F.* Multiple imputation using chained equations for missing data in TIMSS: A case study // Large-Scale Assessments in Education, 2013. Vol. 1. No. 4.

[11] *Acuna E., Rodrigez C.* The treatment of missing values and its effect on classifier accuracy // Classification, clustering and data mining applications. — Berlin–Heidelberg: Springer-Verlag, 2004. P. 639–648.

[12] *Bras L.G., Menezes J.C.* Improving cluster-based missing value estimation of DNA microarray data // Biomol. Eng., 2007. Vol. 24. No. 2. P. 273–282.

[13] *Eskelson B. N. I., Temesgen H., Lemay V., Barrett T. M.,Crookston N. L., Hudak A. T.* The roles of nearest neighbor methods in imputing missing data in forest inventory and monitoring databases // Scand. J. Forest Res., 2009. Vol. 24. No. 3. P. 235–246.

[14] *Pan L.* k-Nearest neighbor based missing data estimation algorithm in wireless sensor networks // Wireless Sensor Network, 2010. Vol. 2. No. 2. P. 115–122.

[15] *Lim J. K., Fuller W. A., Bell W. R.* Variance estimation for nearest neighbor imputation for US Census long form data // Annal. Appl. Stat., 2011. Vol. 5. No. 2A. P. 824–842.

[16] *Jonsson P., Wohlin C.* An evaluation of k-nearest neighbour imputation using Likert data // 10th Symposium (International) on Software Metrics Proceedings, 2004. P. 108–118.

[17] *Kuznetsov M.P., Strijov V.V.* Methods of expert estimations concordance for integral quality estimation // Expert Syst. Appl., 2015. Vol. 41. No. 4. P. 1988–1996.

[18] *Stenina M. M., Kuznetsov M. P., Strijov V. V.* Ordinal classification using Pareto fronts // Expert Syst. Appl., 2015. Vol. 42. No. 14. P. 5947–5953.

[19] *Cooke D. J., Bez H. E.* Computer mathematics. — 1st ed. — Cambridge University Press, 1984. 408 p.

[20] *Wilson D. R., Martinez T. R.* 1997. Improved heterogeneous distance functions // J. Artif. Intell. Res., Vol. 6. P. 1–34.

[21] `http://sourceforge.net/p/mlalgorithms/code/HEAD/tree/Group074/Bakhteev2014MissData/source/.`

References

[1] Horse Colic Data Set. Available at: `https://archive.ics.uci.edu/ml/datasets/Horse+Colic` (accessed June 21, 2015).

[2] Available at: `http://archive.ics.uci.edu/ml/machine-learning-databases/statlog/german/` (accessed June 21, 2015).

[3] Saar-Tsechansky, M., and F. Provost. 2007. Handling missing values when applying classification models. *J. Machine Learning Res. Arch.* 8:1623–1657.

[4] Sharpe, P. K., and R. J. Solly. 1995. Dealing with missing values in neural network-based diagnostic systems. *Neural Comput. Appl.* 3(2):73–77.

[5] Saunders, J. A., N. Morrow-Howell, E. Spitznagel, P. Dori, E. K. Proctor, and R. Pescarino. 2006. Imputing missing data: A comparison of methods for social work researchers. *Social Work Res.* 30(1):19–31.

[6] Batista, G., and M. C. Monard. 2002. A study of k-nearest neighbour as an imputation method. *2nd Conference (International) on Hybrid Intelligent Systems Proceedings.* Santiago, Chile: IOS Press. 251–260.

[7] Durrant, G. B. 2005. Imputation methods for handling item-nonresponse in the social sciences: A methodological review. Available at: http://eprints.ncrm.ac.uk/86/1/MethodsReviewPaperNCRM-002.pdf (accessed October 15, 2015).

[8] Marlin, B. M. 2008. Missing data problems in machine learning. Toronto: University of Toronto. PhD Thesis. 164 p.

[9] Shapfer, J. L. 1999. Multiple imputation: A primer. *Stat. Meth. Medical Res.* 81(1):3–15.

[10] Bouhlila, D. S., and F. Sellaouti. 2013. Multiple imputation using chained equations for missing data in TIMSS: A case study. *Large-Scale Assessments in Education* 1(4).

[11] Acuna, E., and C. Rodrigez. 2004. The treatment of missing values and its effect in the classifier accuracy. *Classification, clustering and data mining applications*. Berlin–Heidelberg: Springer-Verlag. 639–648.

[12] Bras, L. G., and J. C. Menezes. 2007. Improving cluster-based missing value estimation of DNA microarray data. *Biomol. Eng.* 24(2):273–282.

[13] Eskelson, B. N. I., H. Temesgen, V. Lemay, T. M. Barrett, N. L. Crookston, and A. T. Hudak. 2009. The roles of nearest neighbor methods in imputing missing data in forest inventory and monitoring databases. *Scand. J. Forest Res.* 24(3):235–246.

[14] Pan, L. 2010. k-Nearest neighbor based missing data estimation algorithm in wireless sensor networks. *Wireless Sensor Network* 2(2):115–122.

[15] Lim, J. K., W. A. Fuller, and W. R. Bell. 2011. Variance estimation for nearest neighbor imputation for US Census long form data. *Annal. Appl. Stat.* 5(2A):824–842.

[16] Jonsson, P., and C. Wohlin. 2004. An evaluation of k-nearest neighbour imputation using Likert data. *10th Symposium (International) on Software Metrics Proceedings*. 108–118.

[17] Kuznetsov, M. P., and V. V. Strijov. 2015. Methods of expert estimations concordance for integral quality estimation. *Expert Syst. Appl.* 41(4):1988–1996.

[18] Stenina, M. M., M. P. Kuznetsov, and V. V. Strijov. 2015. Ordinal classification using Pareto fronts. *Expert Syst. Appl.* 42(14):5947—5953.

[19] Cooke, D. J., and H. E. Bez. 1984. *Computer mathematics*. 1st ed. Cambridge University Press. 408 p.

[20] Wilson, D. R., and T. R. Martinez. 1997. Improved heterogeneous distance functions. *J. Artif. Intell. Res.* 6:1–34.

[21] Available at: `http://sourceforge.net/p/mlalgorithms/code/HEAD/tree/Group074/Bakhteev2014MissData/source/` (accessed June 21, 2015).