DOI: 10.15514/ISPRAS-2026-38(1)-8

Использование аугментации при обучении на некомплектной выборке

[1,2] *Д.О. Лазарев, ORCID: 0000-0002-6253-6447 <lazarev@ispras.ru>*
[1] *А.В. Шокуров, ORCID: 0000-0002-6801-7728 <shok@ispras.ru>*
[1] *С.А. Фомин, ORCID: 0000-0002-1151-2189 <fomin@ispras.ru>*

[1] *Институт системного программирования им. В.П. Иванникова РАН, Россия, 109004, г. Москва, ул. А. Солженицына, д. 25.*

[2] *Московский физико-технический институт (национальный исследовательский университет), Россия, 141701, Московская область, г. Долгопрудный, Институтский переулок, д.9.*

Аннотация. Исследуется влияние метода аугментации и балансировки некомплектной выборки, содержащей пропущенные значения признаков, на точность прогноза. Результаты для некомплектной выборки сравниваются с результатами для выборки, значения признаков которой полностью заполнены. Предложен новый алгоритм сэмплирования с удалением для аугментации и балансировки некомплектной выборки. В рамках теории вероятностно приближенно корректного (ВПК) обучения авторами была исследована задача обучения на некомплектной выборке. Был оценен рост размерности Вапника-Червоненкиса множества функций при заполнении пропущенных значений фиксированным значением из конечного множества. Было доказано, что требуемый размер выборки для ВПК обучения с достаточной точностью, растет логарифмически медленно с ростом размера этого множества. Установлено, что метод аугментации сэмплированием с удалением, позволяет получить наиболее высокую сбалансированную точность для некомплектных линейно разделимых выборок малого размера. При обучении на выборках среднего и большого размера, во всех рассмотренных случаях, аугментация позволяет получить большее увеличение целевых метрик для некомплектных выборок, чем для полностью заполненных. Таким образом, особенно эффективна аугментация при обучении на некомплектной выборке.

Ключевые слова: машинное обучение; аугментация данных; аугментация табличной выборки; обучение на некомплектной выборке; обучение на выборке малого размера; вероятно приближенно корректное обучение; вложенная кросс-валидация.

Для цитирования: Лазарев Д.О., Шокуров А.В., Фомин С.А. Использование аугментации при обучении на некомплектной выборке. Труды ИСП РАН, том 38, вып. 1, 2026 г., стр. 93–112. DOI: 10.15514/ISPRAS–2026–38(1)–8.

Благодарности: Исследования поддержаны фондом отдела теоретической информатики Института системного программирования им. В.П. Иванникова РАН. Результаты получены с использованием услуг Центра коллективного пользования Института системного программирования им. В.П. Иванникова РАН – ЦКП ИСП РАН.

Data Augmentation for Machine Learning on Missing Data

[1,2] *D.O. Lazarev, ORCID: 0000-0002-6253-6447 <lazarev@ispras.ru>*
[1] *A.V. Shokurov, ORCID: 0000-0002-6801-7728 <shok@ispras.ru>*
[1] *S.A. Fomin, ORCID: 0000-0002-1151-2189 <fomin@ispras.ru>*

[1] *Ivannikov Institute for System Programming of the Russian Academy of Science, 25, Alexander Solzhenitsyn st., Moscow, 109004, Russia.*

[2] *Moscow Institute of Physics and Technology (National Research Institute), 9, Institute alley, Dolgoprudny, Moscow Region, 141701, Russia.*

Abstract. The impact of missing data augmentation and class balancing methods on the prediction accuracy is studied. Results for missing and complete samples are compared. A new algorithm called “drop and replace” for missing data imputation, augmentation and class balancing is proposed. The quality of the “drop and replace” algorithm is compared to the quality of efficient and widely studied augmentation methods, such as SMOTE and mixup. For comparison, 7 samples with sizes from 38 up to 70000 are used. The target metric is balanced accuracy, and the nested cross validation algorithm is used for validation. For the imputation of samples of small sizes, the mean imputation method is used. For the imputation of average or large samples, the MICE method of multiple imputation by chained equations is used.

The problem of training machine learning algorithms on missing data is studied in the probably approximately correct learning setting, or PAC learning. The increase of Vapnik-Chervonenkis dimension, or VC dimension, of the function set due to data imputation with one value from the finite filling set is evaluated. It was shown that the needed sample size for PAC learning with sufficient accuracy, grows logarithmically slow with the increase of the filling set size.

It was proved that “drop and replace” augmentation allows to maximize balanced accuracy when learning on small linearly separable samples containing missing values. For larger samples, SMOTE and mixup augmentation methods are most effective. For 8 samples out of 11, data augmentation, which increments sample size, increases classification accuracy; in 2 of 3 other cases, sample balancing alone can increment classification accuracy; in the final case, the initial sample is balanced and augmentation cannot increase classification accuracy. When learning on average and large-sized samples, in all 3 cases, augmentation more significantly improves accuracy for missing datasets, then for complete ones. So, augmentation is more effective when learning on missing data.

Keywords: machine learning; data augmentation; tabular data; machine learning on missing data; learning on small sized dataset; probably approximately correct learning; nested cross validation.

For citation: Lazarev D.O., Shokurov A.V., Fomin S.A. Data augmentation for machine learning on missing data. Trudy ISP RAN/Proc. ISP RAS, vol. 38, issue 1, 2026, pp. 93-112 (in Russian). DOI: 10.15514/ISPRAS-2025-38(1)-8.

Acknowledgements. The research was supported by the fund of the Department of Theoretical Computer Science of Ivannikov Institute for System Programming of the RAS. The results were obtained using the services of the Shared Use Center of the Ivannikov Institute for System Programming of the PAS – the SUC ISP RAS.

1. Введение

Обучение на некомплектной выборке, или на выборке, содержащей незаполненные значения некоторых признаков, требуется во многих реальных приложениях, включая медицину и биологию [1-2]. Наличие пропущенных значений в выборке может быть обусловлено разными причинами, например, неправильным функционированием измерительных приборов, отсевом участников в продолжение наблюдения или нежеланием участников исследования отвечать на вопрос, например, о своем доходе.

Сформулируем постановку задачи обучения на некомплектной выборке.

Построить алгоритм машинного обучения, который позволяет строить прогноз класса для некомплектных объектов. Требуется выбрать способ валидации, минимизирующий

переобучение, и получить несмещенную оценку точности прогноза для выборки малого размера. При обучении, восстановить пропущенные значения и сделать аугментацию обучающей выборки. Исследовать влияние способа аугментации выборки и метода заполнения пропущенных значений на точность прогноза. При исследовании, будут использоваться несколько выборок.

В работе [1] приводится следующая классификация некомплектных выборок по вероятности отсутствия значения признака:

- Пропущенные значения полностью случайны (missing completely at random, MCAR). Вероятность отсутствия каждого значения признака не зависит от этого значения и от класса, которому соответствующий объект принадлежит.
- Пропущенные значения случайны (missing at random, MAR). Вероятность отсутствия каждого значения не зависит от самого этого значения, но может зависеть от класса, которому объект принадлежит.
- Пропущенные значения неслучайны (missing not at random, MNAR). Вероятность того, что значение пропущено, может зависеть от этого значения.

Отметим, что во всех случаях, включая MCAR и MAR, вероятности отсутствия значения для любых двух различных признаков одной выборки могут различаться.

Метод вложенной кросс-валидации, используемый для валидации модели машинного обучения, позволяет получать несмещенную оценку точности для любого типа некомплектной выборки. Но при восстановлении пропущенных значений предполагается, что некомплектная выборка принадлежат одному из типов MCAR или MAR.

Таким образом, в настоящей работе рассмотрены некомплектные выборки, являющиеся либо MAR, либо MCAR выборками. Исследуем выборки следующих типов: 2 некомплектных выборки [3, 4] типа MAR, для которых вероятность отсутствия признака зависит от класса и 5 полностью заполненных выборок [5, 6, 7, 8, 9], которые преобразуем в некомплектные MCAR-выборки случайным удалением значений признаков.

При восстановлении, пропущенные значения признаков заполняются неточно. Поэтому, в настоящей работе предложена аугментация выборки с восстановлением пропущенных значений. Таким образом, восстанавливаемые пропущенные значения из обучающей выборки, заполняются не одним фиксированным значением, а целым множеством значений из вероятностного распределения, построенного на основе заполненных значений соответствующего признака.

Похожий подход применялся в работе [10], где адаптивная аугментация некомплектной выборки с использованием генеративно-состязательных сетей позволяла повысить точность прогноза в случае, когда тестовая и обучающая выборки имели различные распределения пропущенных значений.

Наиболее часто аугментация применяется при обучении на таких типах многомерных выборок, как изображения [11], аудиоданные [12] или для выборок, возникающих при обучении больших языковых моделей [13]. Однако, многие методы аугментации данных, применяемые преимущественно для многомерных выборок, такие, например, как метод mixup [14] “перемешивания” с помощью построения выпуклых комбинаций из пар объектов, могут также применяться при обучении на табличных выборках. Метод SMOTE [15], наиболее широко применяемый для балансировки выборки, т.е. для того, чтобы сделать равными размеры классов в выборке, также может быть использован для аугментации табличной выборки. Все рассматриваемые в настоящей работе методы аугментации данных, также позволяют балансировать выборку.

В настоящей работе предложен метод сэмплирования с удалением для аугментации и балансировки выборки. Наиболее высокую точность данный метод аугментации позволяет получить при обучении на некомплектной линейно разделимой выборке малого размера.

Предлагаемый подход аугментации данных для обучения на некомплектной выборке является новым методом, который ранее в литературе не встречался.

Также исследовалось влияние аугментации некомплектной выборки на точность прогноза. Результаты для некомплектной выборки сравнивались со случаем полностью заполненной выборки. Для 3 из 4 некомплектных выборок размера более 500, аугментация позволяла улучшить точность прогноза. Для оставшейся выборки PPMI, было достаточно балансировки выборки для увеличения точности. Для полностью заполненных выборок, аугментация позволяла увеличить точность прогноза в 2 случаях из 4, а увеличение точности при этом наблюдалось менее существенное, чем для некомплектных выборок. Таким образом, проведенные исследования свидетельствуют о том, что для выборок достаточно большого размера, аугментация позволяет увеличить точность прогноза для некомплектных выборок.

В следующем разделе, приведем формальную постановку решаемой задачи в рамках теории [16] Валианта вероятностно приближенно корректного обучения (ВПК обучения) и теоретически исследуем ВПК-обучаемость задачи обучения на некомплектной выборке.

2. Исследование и постановка задача обучения на некомплектной выборке в рамках теории ВПК обучения

Теория вероятностно приближенно корректного (ВПК) обучения [16, 17] позволяет теоретически исследовать алгоритмы машинного обучения. В рамках классической задачи обучения с учителем, имея выборку $\{(x_i, y_i)_{i=1}^m\}, x_i \in X, y_i \in Y$, теория ВПК обучения позволяет выбрать множество $\mathcal{H}$ гипотез, которому должна принадлежать обобщающая функция $h: X \to Y$, являющаяся результатом обучения, таким образом, чтобы одновременно уменьшить ошибку аппроксимации и избежать переобучения.

Элементы теории ВПК обучения, включающие определения агностической ВПК-обучаемости, VC-размерности Вапника и Червоненкиса и лемму Зауэра-Шелаха, приведены в приложении 1.

Сформулируем задачу обучения на некомплектной выборке в терминах теории ВПК обучения [17]. Пространство признаков X_i для каждого признака i дополним значением $\perp$, означающим то, что данный признак пропущен. Соответственно, множество гипотез определим на множестве $X^\perp = \prod_{i=1}^n (X_i \cup \perp)$ некомплектных выборок.

Дадим общую постановку, взяв за основу некоторое множество гипотез $\mathcal{H}$, определенное только на полностью заполненных объектах из X. При этом, заполненные значения исходной выборки могут содержать случайный шум. Аналогично [18], будем использовать алгоритм $Fill: X^\perp \to X$ для восстановления пропущенных значений.

Приведем определение агностического ВПК обучения на некомплектной выборке и оценим сверху необходимый размер выборки для ВПК обучения с заданными параметрами вероятности и аппроксимации.

Пусть $X = \prod_{i=1}^n X_i$ – исходное пространство объектов (считаем значения признаков X_i вещественными), а $X^\perp = \prod_{i=1}^n (X_i \cup \perp)$ – пространство некомплектных выборок. Пусть также задано множество $CFill$ функций $Fill: X^\perp \to X$ восстановления пропущенных значений. Пусть $\mathcal{H}$ – множество гипотез, являющееся классом характеристических функций $h: X \to \{0,1\}$.

Определение ВПК обучения на некомплектной выборке требует, чтобы множество $CFill$ функций для восстановления пропущенных значений было задано. Определим агностическое обучение на некомплектной выборке $X^\perp$, используя обучение на полностью заполненной выборке X для множества гипотез $\mathcal{H}$, являющегося классом характеристических функций $h: X \to \{0,1\}$. Определим на некомплектной выборке $X^\perp$ класс гипотез $\mathcal{H}'(X^\perp) = (\mathcal{H}(CFill(X^\perp))): X^\perp \to \{0,1\}$ как композицию восстановления пропущенных значений

алгоритмов $Fill$ из класса $CFill$ и применения гипотезы из множества $\mathcal{H}$ к комплектному объекту из множества X.

Определение (агностическое ВПК обучение на некомплектной выборке). Говорим, что множество $(\mathcal{H}, CFill)$ – агностически ВПК-обучаемое, тогда и только тогда, когда множество гипотез $\mathcal{H}'(X^{\perp}) = (\mathcal{H}(CFill(X^{\perp}))) \colon X^{\perp} \to \{0,1\}$, получаемых в результате применения к некомплектным объектам сначала одной функции $Fill$ для восстановления пропущенных значений из множества $CFill$, а затем – некоторой характеристической функции $h \in \mathcal{H}$, является агностически ВПК-обучаемым.

Отметим, что бинарные метки объектов $y_i \in \{0,1\}$ – полностью заполнены для всех i.

Следующая фундаментальная теорема ВПК обучения дает критерий обучаемости для некомплектной выборки.

Теорема 1 (Фундаментальная теорема ВПК обучения). Для некомплектной выборки, множество $(\mathcal{H}, CFill)$ – агностически ВПК-обучаемое тогда и только тогда, когда $VCdim(\mathcal{H}'(\mathcal{H}, CFill)) < \infty$.

Приведем полученные в настоящей работе новые теоретические результаты, связанные с VC-размерностью некомплектной выборки при заполнении пропущенных значений фиксированными элементами из некоторого, заранее известного множества.

Лемма 1 (Агностическая ВПК-обучаемость на некомплектной выборке). Пусть $CFill_{const}$ – множество методов $Fill \colon X^{\perp} \to X$ восстановления пропущенных значений, представляет из себя множество методов заполнения константой из множества $X_{fill} = \{x_{fill}^{j}\}_{j=1}^{k}, x_{fill}^{j} \in X \ \forall 1 \le j \le k$ конечного размера k. Пусть $\mathcal{H}$ – множество гипотез $h \colon X \to \{0,1\}$, имеющее конечную VC-размерность: $VCdim(\mathcal{H}) = d$.

Тогда выполняется следующее:

1. При $|X_{fill}| = 1$, $VCdim(\mathcal{H}'(\mathcal{H}, CFill_{const})) \le d$.
2. При $|X_{fill}| = k > 1$, $VCdim(\mathcal{H}'(\mathcal{H}, CFill_{const})) < 2d\ log_2\ (k^{1/d} + 31)$.

(В лемме, $\mathcal{H}'(\mathcal{H}, CFill_{const}) = \mathcal{H}(CFill_{const}(X^{\perp}))$ – множество гипотез, получаемых в результате применения к некомплектным объектам сначала одного метода $Fill$ для восстановления пропущенных значений из множества $CFill_{const}$, а затем – некоторой характеристической функции $h \in \mathcal{H}$).

Лемма 1 доказана в приложении 2.

Практическая значимость леммы 1 обусловлена медленным ростом мультипликатора $2\ log_2\ (k^{1/d} + 31)$ в неравенстве $VCdim(\mathcal{H}'(\mathcal{H}, CFill_{const})) < 2d\ log_2\ (k^{1/d} + 31)$ в зависимости от размера k множества X_{fill}. При заполнении каждого из n_f пропущенных значений множеством точек из сетки, являющейся декартовым произведением n_f отрезков, содержащих N точек, мультипликатор $2\ log_2\ (k^{1/d} + 31)$ в ограничении сверху у $VCdim(\mathcal{H}'(\mathcal{H}, CFill_{const}))$ растет со скоростью $O(\frac{n_f}{d} log\ N)$, т.к. $k = N^{n_f}$.

Согласно количественной формулировке фундаментальной теоремы ВПК обучения [19], для множества гипотез $\mathcal{H} \colon VCdim(\mathcal{H}) = d < \infty$, существуют такие константы $c_1 > c_0 > 1$, что для случайной выборки размера не менее $m(\epsilon, \delta)$, удовлетворяющего неравенству $c_0 \frac{d + ln(1/\delta)}{\epsilon^2} < m(\epsilon, \delta) < c_1 \frac{d + ln(1/\delta)}{\epsilon^2}$, выполняется условие агностического ВПК обучения: с вероятностью не менее $1 - \delta$, минимизирующий эмпирический риск ВПК-обучатель, возвращает такую гипотезу $h \in \mathcal{H}$, для которой выполняется: $L_D(h) \le L_D(\mathcal{H}) + \epsilon$.

Соответственно, при обучении на некомплектной выборке, при котором сначала заполняем пропущенное значение одной из N^{n_f} констант из сетки, требуемый размер $m(\epsilon, \delta)$ для обучения с заданной точностью увеличится в $\frac{n_f}{d} ln\ N$ раз по сравнению со случаем полностью заполненной выборки при выборе точки для заполнении пропущенных значений по сетке и

при выборе гипотезы $h \in \mathcal{H}$. Иллюстрация заполнения пропущенных значений элементами конечного множества приведена на рис. 1.

Рис. 1. Заполнение пропущенных значений элементом множества X_{fill} при обучении на некомплектной выборке.

Fig. 1. Missing data imputation with an element of finite set X_{fill} when learning to solve classification task.

2.1 Формальная постановка задачи обучения на некомплектной выборке

Приведем формальную постановку решаемой задачи обучения на некомплектной выборке и исследования влияния аугментации выборки на точность прогноза. Дана выборка $S = \{(x_i, y_i)\}_{i=1}^{n}$, являющаяся объединением численных представлений объектов $x_i \in X^{\perp}, i \in \{1, \dots, n\}$ и ответов $y_i \in Y, i \in \{1, \dots, n\}$, принадлежащих конечному множеству Y. Все элементы выборки S выбраны независимо в совокупности согласно некоторому неизвестному распределению D. Объекты могут быть некомплектными, т.е. часть значений признаков может быть пропущена. Полностью заполненные ответы y_i позволяют определить, к какому классу относится объект.

Целью работы является:

- Построение функции прогноза класса $\hat{y}(x, S)$, которая для каждого, возможно – некомплектного объекта $x \in X^{\perp}$, строит прогноз ответа. При этом стремимся минимизировать ожидаемую ошибку $E_{(x,y)\sim D}(l(\hat{y}(x, S), y)) \to min$ для метрики $l: Y^2 \to \mathbb{R}$.
- Выбор целевой функции и метода валидации для получения несмещенной оценки точности прогноза с малой дисперсией.
- Разработка и сравнение методов аугментации некомплектной выборки, использующих методы восстановления пропущенных значений и позволяющих получать высокую точность прогноза при обучении на некомплектной выборке алгоритма машинного обучения. Сравнение результатов со случаем полностью заполненной выборки.

3. Обзор существующих решений

Приведем обзор существующих решений и сравним их с решением, предложенным в настоящей работе.

В работе [20], в рамках фреймворка VIME, предложен метод аугментации табличной выборки, позволяющий получить высокую точность прогноза. Метод аугментации также может быть применен для восстановления пропущенных значений. Идея искажения выборки с последующим ее восстановлением из фреймворка VIME была использована в настоящей работе при построении метода аугментации сэмплирования с удалением. Однако, в отличие

от настоящей работы, где исследуется обучение с учителем, фреймворк VIME предназначен для использования при обучении с частичным подкреплением или при обучении без учителя. При обучении с частичным подкреплением, при обучении используются как размеченные, так и неразмеченные объекты выборки, а при обучении без учителя, все объекты выборки не размечены.

В работе [21], предложен автокодировщик с шумоподавлением для восстановления пропущенных значений. Такой автокодировщик обучен восстанавливать пропущенные значения путем подачи ему на вход зашумленной выборки.

Автокодировщики позволяют получить более высокую точность прогноза при наличии зависимости вероятности пропуска значения выборки от данного значения [22]. Таким образом, в случае MNAR, автокодировщики позволяют более точно восстанавливать пропущенные значения по сравнению с другими методами восстановления пропущенных значений. В случаях MCAR и MAR, метод восстановления пропущенных значений максимизацией эмпирического правдоподобия, превосходит автокодировщик по точности прогноза.

Точность обучения при использовании автокодировщика с шумоподавлением для восстановления пропущенных значений сравнивалась в [21] с точностью при использовании метода MICE цепных уравнений [23]. Для некомплектной выборки типа MNAR, восстановление пропущенных значений автокодировщиком с подавлением шума, позволяет получить более высокую точность прогноза, чем восстановление пропущенных значений методом MICE цепных уравнений.

В отличие от работы [21], в настоящей работе используются некомплектные выборки типов MCAR и MAR. Для восстановления пропущенных значений используется метод MICE цепных уравнений. Сравниваются между собой в настоящей работе не методы восстановления пропущенных значений, а методы аугментации выборки. Также, точность обученной модели после восстановлении пропущенных значений, сравнивается со случаем обучения на некомплектной выборке.

В настоящей работе, также исследована ВПК-обучаемость в случае обучения на некомплектной выборке. Аналогичная задача исследовалась ранее, например в [18, 24]. Отличие предложенной в настоящей работе постановки заключается в том, что в [18] исследуется задача ВПК обучения, а не агностического ВПК обучения. Класс функций $\mathcal{H}$ представлен в [18] булевыми формулами. В работе [24] делается предположение о возможности сколь угодно точного восстановления пропущенных значений для выборки достаточно большого размера. Таким образом, в моделях [18, 24] случай зашумленной некомплектной выборки не рассматривается, что ограничивает практическое применение полученных результатов. В настоящей работе, исследуется агностическая ВПК обучаемость на некомплектной выборке, что позволяет моделировать обучение на зашумленной некомплектной выборке.

4. Методы обучения и валидации на некомплектной выборке

При обучении моделей машинного обучения на некомплектной выборке, выделим следующие этапы:

1. Выбор метода обучения на некомплектной выборке
2. Восстановление пропущенных значений в обучающей и в тестовой выборках
3. Выбор целевой метрики для оценки качества прогноза
4. Выбор схемы валидации модели
5. Выбор метода аугментации выборки

На начальном этапе, нужно сделать выбор метода обучения: будем ли обучаться на некомплектной выборке, или предварительно обучим алгоритм восстановления

пропущенных значений. Выбрав восстановление пропущенных значений, предварительно, будем удалять из выборки объекты, у которых пропущено слишком много значений, т.к. точность прогноза данных объектов может быть значительно ниже среднего.

Выбор целевой метрики должен учитывать несбалансированность классов выборок. Выбор метода валидации модели машинного обучения, особенно важен для выборок малого размера, для которых стандартная схема разбиения исходной выборки на обучающую и тестовую выборки, дает слишком большую дисперсию полученной оценки точности прогноза.

4.1 Обучение на некомплектной выборке

Согласно [2], для некомплектной выборки, может быть использован один из следующих подходов для обучения ML-модели:

- Удаление некомплектных объектов, или тех объектов, у которых пропущены некоторые значения признаков. Данный подход может привести к удалению значительного числа объектов из выборки. Позволяет получить несмещенную оценку точности только в случае, когда выборка относится к типу MCAR или MAR. В случае MAR нужно сделать балансировку выборки, т.к. после удаления некомплектных объектов, баланс классов может измениться.
- Обучение ML-модели, предназначенной для работы с некомплектной выборкой. К таким моделям относятся алгоритмы машинного обучения, использующие лишь попарные корреляции между признаками и модель CatBoost [25].
- Восстановление пропущенных значений (англ. imputation).

Часть алгоритмов машинного обучения не предназначена для работы на некомплектных выборках, другие алгоритмы могут показывать значительное снижение точности при обучении на некомплектной выборке по сравнению с обучением на выборке, пропущенные значения в которой восстановлены.

Поэтому, в настоящей работе, применялись первый и последний подходы. Из обучающей и из тестовой выборок удалялись объекты, у которых пропущено слишком большое число наиболее значимых признаков. Выбор наиболее значимых признаков для обучения с использованием t-критерия Уэлча позволял уменьшить количество удаляемых объектов. Затем пропущенные значения в выборке восстанавливались.

Результаты сравнивались со случаем обучения на некомплектной выборке без предварительного восстановления пропущенных значений. Для рассмотренной выборки, точность прогноза снизилась при обучении на некомплектной выборки по сравнению со случаем восстановления пропущенных значений перед обучением ML-модели.

4.2 Восстановление пропущенных значений

Разделим, согласно [2], методы восстановления пропущенных значений в выборке на 2 типа:

- Одномерные методы. К данным методам относится восстановление пропущенного значения либо средним медианным значением для непрерывных признаков, или модой для категориальных. Восстановление пропущенного значения признака происходит с использованием имеющихся в исходной выборке значений этого же признака признаков у других объектов.
- Многомерные методы восстановления пропущенных значений. Существуют различные методы многомерного восстановления: метод k ближайших соседей KNN [26], восстановление с использованием метода нечеткой кластеризации C-средних FCM [27], восстановление с использованием байесовского метода главных компонент [28], восстановление методом MICE цепных уравнений [23] и др. При

восстановлении пропущенных значений одним из данных методов, учитываются значения заполненных признаков в исходной выборке у того объекта, пропущенные значения которого восстанавливаются.

В [29] сравнивались 6 методов восстановления пропущенных значений, включая одномерный метод заполнением средним и многомерные подходы: метод k ближайших соседей, метод FCM нечеткой кластеризации C-средними и байесовский метод главных компонент. Наиболее высокая точность достигалась у двух последних методов, причем метод FCM позволял получить наибольшую точность при восстановлении значений выборок малого размера.

В работе [30], исследовались методы восстановления пропущенных значений некомплектной выборки типа MNAR. Сравнивались методы восстановления пропущенных значений средним, MICE цепных уравнений, k ближайших соседей, SoftImpute. Наиболее высокую точность восстановления пропущенных значений позволяет получить MICE, а метод k ближайших соседей позволяет получить наиболее высокую точность классификации. В [31], к сравнению добавлено восстановление пропущенных значений с использованием генеративного искусственного интеллекта. Последние показывали лучшую точность при большей доле пропущенных значений в выборке (60% и выше), тогда как при относительно малой доле пропущенных значений (40% и менее), наибольшая точность восстановления пропущенных значений достигается при использовании метода MICE цепных уравнений. Согласно проведенным в работе [32] вычислительным экспериментам, наибольшую точность восстановления пропущенных значений позволяют получить различные модификации метода K ближайших соседей.

В настоящей работе, сравнивались методы восстановления пропущенных значений средним, метод MICE цепных уравнений, метод нечеткой кластеризации C-средних FCM, метод K ближайших соседей и метод Miss Forest [33]. Для выборок малого размера, содержащих до 50 объектов, наибольшую точность классификации позволял получить метод восстановления пропущенных значений средним; а для выборок большего размера, содержащих более 500 объектов в обучающей выборке, наиболее эффективен метод MICE цепных уравнений. Поэтому методы восстановления пропущенных значений средним и методом MICE цепных уравнений, и были использованы в настоящей работе.

Отметим, что имеется различие при восстановлении пропущенных значений в обучающей и в тестовой выборке. В первом случае, класс, к которому принадлежит объект, можем использовать при восстановлении пропущенных значений; в случае восстановления пропущенных значений тестовой выборки, класс объекта не может быть использован, в противном случае возникает проблема утечки данных.

4.3 Используемые метрики

Все выборки, кроме MNIST, являются несбалансированными. Наибольший дисбаланс классов наблюдается для выборки PPMI [3], где 91.7% объектов принадлежат одному классу из двух. Поэтому, стандартной точности $accuracy$ недостаточно для оценки точности прогноза на несбалансированной выборке [10]. Для преодоления этого недостатка будем использовать метрики сбалансированная точность $balanced\,accuracy$ и коэффициент корреляции Мэтьюса [34] для оценки точности прогноза. Сбалансированная точность равна точности $accuracy$ в случае сбалансированной выборки.

4.4 Выбор метода валидации модели

При использовании схемы кросс-валидации при обучении с подкреплением на выборке малого размера, из-за переобучения полученная точность может падать с ростом размера исследуемой выборки [35]. Происходит это из-за переобучения, т.к. исходная выборка

используется как для обучения модели, так и для выбора ее гиперпараметров. Поэтому на выборках малого размера, очень важен выбор схемы валидации, позволяющий получить несмещенную оценку точности, минимизируя при этом переобучение.

Валидация ML-модели на отложенной тестовой выборке, позволяет получить несмещенную оценку точности. Однако, при этом, для выборки малого размера, получается большая дисперсия оценки точности, т.к. на точность прогноза будет существенно влиять выбор тестовой выборки. Таким образом, валидацию на отложенной тестовой выборке будем применять лишь для выборок большого размера, содержащих более 10000 объектов.

Согласно [36], для получения несмещенной оценки точности с малой дисперсией, следует использовать метод вложенной кросс-валидации [37] для сравнения методов аугментации выборок малого размера. Будем использовать этот метод для всех выборок малого и среднего размера, содержащих менее 2000 объектов до аугментации.

В схеме вложенной кросс-валидации, выделяются следующие шаги:

1. Выбор обучающей и тестовой выборок. Элементы обучающей выборки выбираются из исходной случайным сэмплированием некоторого числа объектов без повторения. В тестовую выборку входят все объекты, не вошедшие в обучающую. В настоящей работе, размер обучающей и тестовой выборки на каждом шаге вложенной кросс-валидации составляет около 90% и 10% от размера исходной выборки соответственно.
2. Затем, обучающая выборка балансируется и аугментируется одним из 5 методов, качество которых сравнивается между собой и со случаем отсутствия аугментации и балансировки выборки. Далее, на обучающей выборке после аугментации обучается стандартизатор, который будет применяться как при обучении, так и при прогнозе на тестовой выборке. На выбранных признаках обучается модель машинного обучения, например, машина опорных векторов (SVM) или градиентный бустинг. Гиперпараметры модели выбираются с использованием кросс-валидации на обучающей выборке. Данная кросс-валидация называется внутренней кросс-валидацией в схеме вложенной кросс-валидации.

В дальнейшем, шагом вложенной кросс-валидации, будем называть одно повторение шагов 1, 2.

3. На данном шаге выполняется внешняя кросс-валидация. Шаги 1, 2 повторяются большое число раз (в проведенных вычислительных экспериментах, число повторений составляет 200 для малых выборок размером менее 50 объектов и 20 раз – на больших выборках, размером более 500). Каждый раз происходит обучение для нового случайно выбранного обучающего набора объектов. Точность прогноза класса оценивается на объединении всех тестовых выборок, полученных на шаге внутренней кросс-валидации. Некомплектные объекты со слишком большим числом пропущенных значений из выборки удаляются, а пропущенные значения для оставшихся объектов восстанавливаются. Итоговая точность оценивается на объединении прогнозов по всем тестовым выборам. При этом, каждый объект с достаточным количеством заполненных значений признаков, входит в это объединение прогнозов одинаковое количество раз.

Диаграмма схемы вложенной кросс-валидации, используемой в настоящей работе, приведена на рис. 2.

При обучении на 5 выборках, размер которых менее 2000, для оценки точности прогноза ML-модели, в настоящей работе используем схему вложенной кросс-валидации. Для оставшихся 2 выборок, размер которых превосходит 10000, используем независимую тестовую выборку для валидации. Согласно [38], обе данные схемы валидации позволяют получить несмещенные оценки точности. Для выборки малого размера, оценка точности на

независимой тестовой выборке может быть неточной из-за сильной зашумленности получаемой таким образом оценки. Поэтому, при обучении на малых выборках, важно использовать схему вложенной кросс-валидации.

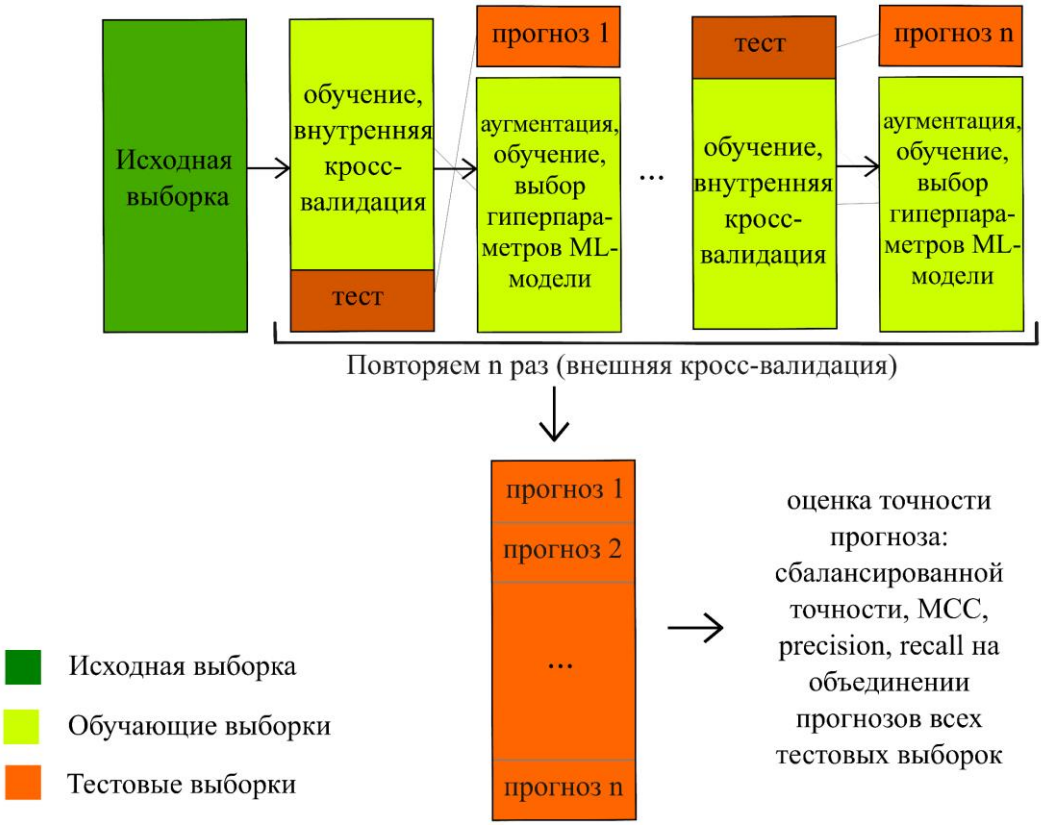

Рис. 2. Схема вложенной кросс-валидации.
Fig. 2. Nested cross validation scheme.

4.5 Аугментация выборки

Сравним предложенный в настоящей работе метод аугментации сэмплированием с удалением и другие методы аугментации и балансировки выборки. Эти методы аугментации можно разделить на 2 категории: во-первых, широко применяемые используемые ранее методы аугментации SMOTE и mixup, и, во-вторых, простые методы, включающие в себя минимально изменяющий выборку подход сэмплирования без удаления и строящий случайные комбинации значений признаков каждого класса метод случайного сэмплирования. Ниже подробно опишем все методы аугментации выборок, которые будем сравнивать между собой и со случаем отсутствия аугментации:

1. **Аугментация сэмплированием без удаления.** Строим синтетическую выборку случайным сэмплированием из исходной выборки нужного числа объектов, у которых заполнено достаточное количество значений признаков. При наличии пропущенных значений в выборке, используется метод MICE цепных уравнений, или одномерный метод восстановления пропущенных значений средним значением. К восстановленному значению признака добавляется также случайный шум с меньшей дисперсией, чем дисперсия значения соответствующего признака в исходной некомплектной выборке. Особенность подхода заключается в том, что, за

исключением восстановления пропущенных значений, в выборку не добавляются новые синтетические значения, а вместе с ними и новые связи между значениями признаков.

2. **Аугментация сэмплированием с удалением.** Сначала также, как и при аугментации сэмплированием без удаления, сэмплируем нужное число объектов из имеющейся, возможно некомплектной, выборки. Далее удаляем каждое заполненное значение признака с вероятностью $\frac{1}{2}$, после чего восстанавливаем все пропущенные значения. Преимуществом данного подхода является как сохранение части заполненных значений некомплектной выборки, так и добавление новых, зашумленных значений. Последнее можно рассматривать как регуляризацию модели машинного обучения [39] на уровне выборки.

Похожий подход, заключающийся в искажении и последующем восстановлении выборки, применяется и во фреймворке VIME [20] обучения со слабым подкреплением.

3. **Аугментация случайным сэмплированием**, или случайным выбором значений признаков. Простейший метод аугментации, где новые объекты строим случайным сэмплированием из набора значений каждого признака для каждого класса исходной выборки. При данном методе аугментации, значение каждого признака синтетического объекта выбирается из соответствующего класса исходной выборки случайным образом. Недостаток такого подхода заключается в том, что могут теряться важные для прогноза зависимости и связи между значениями признаков. Метод может давать высокую точность прогноза в случае малой корреляции между значениями различных признаков внутри классов.

Отличие подхода от аугментации сэмплированием без удаления заключается в сэмплировании отдельных значений признаков, а не целых объектов.

4. **Использование метода SMOTE** [15] построения синтетических объектов с использованием метода k ближайших соседей. При восстановлении пропущенных значений, используем метод цепных уравнений MICE или одномерный метод восстановления пропущенных значений средним.
5. **Метод mixup** [14] аугментации**.** Заключается в создании новых синтетических объектов выпуклой комбинацией пар уже имеющихся объектов внутри каждого из классов.

Для каждого из методов аугментации выборки, помимо аугментации, также производится балансировка выборки, в результате которой размеры всех классов после аугментации имеют одинаковый размер.

На рис. 3., визуализированы различные методы аугментации данных полностью заполненной выборке MNIST, содержащей написанные от руки изображения цифр.

5. Описание выборок

Опишем 7 выборок, которые будем использовать для сравнения различных методов аугментации. 6 выборок из 7 являются медицинскими выборками. На практике, данный тип выборок часто содержит некомплектные объекты. Каждая из 7 выборок содержит хотя бы один категориальный признак, т.к. целевой класс, к которому принадлежит объект, является категориальным.

1. ***Prodromal PD.*** Выборка [4] для прогноза болезни Паркинсона на продромальной стадии по биомаркерам в гуморальных средах. Объект в выборке – пациент, признаки – его параметры крови. Доля пропущенных значений – 23.7%. Выборка подвергалась предварительной обработке – удалению тех признаков, которые заполнены менее, чем у половины объектов.

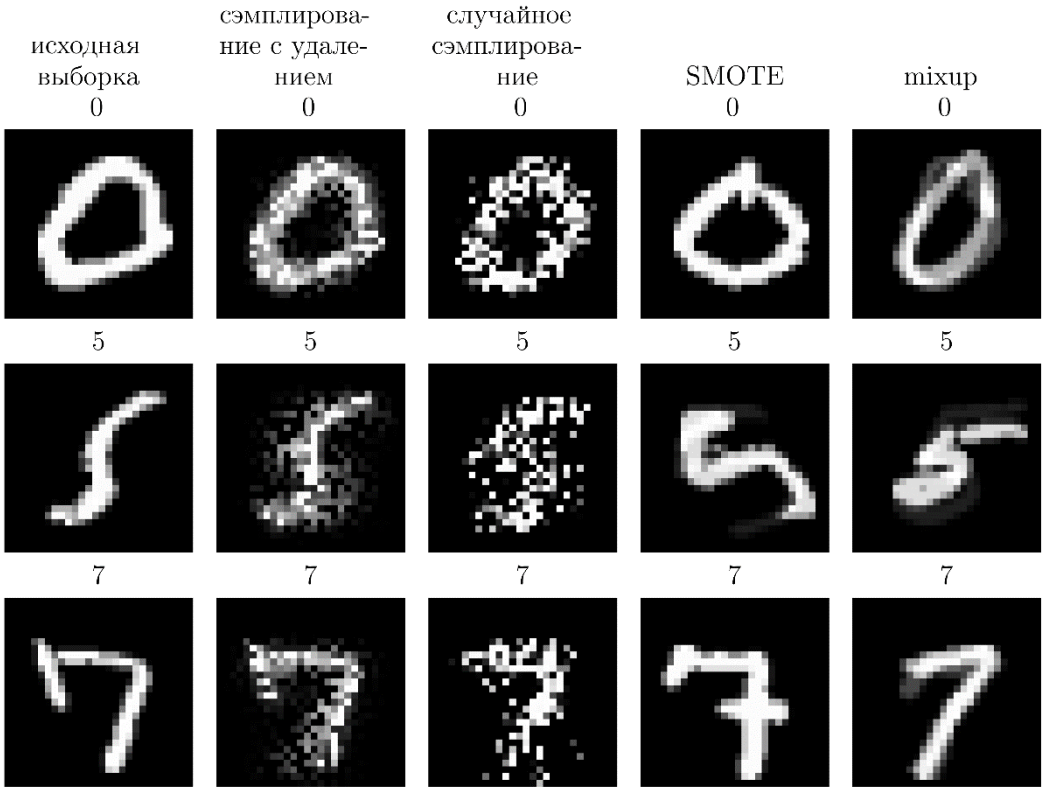

Рис. 3. Визуализация различных методов аугментации данных с использованием выборки MNIST.
Fig. 3. Augmentation method comparison on MNIST dataset.

2. ***Acute Leukemia.*** В выборке [5] собраны почти 7000 генов у 72 пациентов. Пациенты делятся на две группы: пациенты с острым лимфобластным лейкозом и острым миелобластным лейкозом. Выборка включает в себя обучающую выборку размером 38 и тестовую размером 34. Тестовая и обучающая выборка получены в разных лабораториях, исследования проведены по различным гуморальным средам. Также, тестовая выборка включала пациентов детского возраста, а обучающая не включала. Поэтому используем только обучающую выборку, из которой оставим только 50 наиболее значимых генов согласно t-тесту Уэлча.
3. ***PPMI***. Инициативой PPMI поиска маркеров прогрессирования БП [3], была собрана выборка, содержащая более 300 здоровых пациентов из контрольной группы и более 3300 пациентов с диагнозом БП. Выборка содержит демографические данные, снимки ПЭТ-томографии, параметры крови, биомаркеры в спинномозговой жидкости и другие признаки.
4. ***MNIST***. Выборка MNIST [6] содержит нормированные рукописные изображения цифр от 0 до 9 размером 28×28 пикселей. Таким образом, данные содержат 784 признаков. В обучающей выборке содержится 60000 изображений, размер тестовой выборки – 10000. Изображения в выборке MNIST центрированы, т.е. центр масс каждого изображения совпадает с центром изображения.
5. ***EyeState***. В выборке [8] значения признаков получены с использованием ЭЭГ. Объекты принадлежат двум классам: для первого закрыты глаза при измерении ЭЭГ, для второго – закрыты.

6. ***Breast Cancer Wisconsin***. Выборка [9] содержит цифровизированные значения, полученные из снимков тонкоигольной аспирационной биопсии опухоли молочной железы. Выборка содержит доброкачественные и злокачественные опухоли.
7. ***Cardiotocography***. Выборка [7] содержит измерения кардиотокограммы плодов, разделенные на 3 класса опытными акушерами-гинекологами: нормальный результат, диагноз под вопросом и наличие патологии.

Табл. 1. Описание использованных в настоящей работе выборок.
Table 1. Sample description.

Название выборки	Число объектов в выборке	Доля пропущенных значений	Число признаков	Число категориальных признаков	Число классов	Доля максимального класса
Prodromal PD	45	23.7%	11	1	2	57.8%
Acute Leukemia	38	0%	51	1	2	71.1%
PPMI	3625	3.8%	35	5	2	91.7%
MNIST	70000	0%	785	1	10	10%
EyeState	14980	0%	15	1	2	55.1%
Breast Cancer Wisconsin	569	0%	31	1	2	62.7%
Cardiotocography	2126	0%	42	12	3	77.8%

Первые 2 выборки размером менее 50, будем называть малыми, MNIST и EyeState, содержащие более 10000 объектов назовем большими, а оставшиеся 3 выборки размером от 500 до 4000, назовем средними.

Заметим, что из 7 выборок только 2 содержат пропущенные значения и являются некомплектными. Выборки PPMI и prodromal PD имеют тип MAR, т.к. в контрольной группе у них больше доля некомплектных значений, чем в группе риска. 5 полностью заполненных выборок, за исключением MNIST, будем преобразовывать в некомплектные выборки типа MCAR удалением части значений признаков из выборки случайным образом.

В 5 из 7 выборок, задача классификации бинарная; в выборке “Cardiotocography”, количество классов равняется 3, выборка MNIST содержит 10 классов изображений. Все выборки, за исключением prodromal PD и PPMI, находятся в открытом доступе.

6. Результаты вычислительных экспериментов

В настоящем разделе приведем результаты следующих вычислительных экспериментов:

- Сравним точность классификации в случае восстановления пропущенных значений и без него.
- Выясним, насколько аугментация позволяет увеличить точность классификации, сравним разные методы аугментации выборки и найдем наиболее эффективные.

Выберем критерии сравнения точности классификации. Пусть для двух сравниваемых измерений получены оценки α_1 и α_2 сбалансированной точности. Во-первых, будем использовать увеличение сбалансированной точности $\alpha_2 - \alpha_1$. Однако, одно изменение сбалансированной точности не позволяет объективно оценить точность прогноза, т.к. увеличение точности на одинаковую величину при малом α_1 (при качестве прогноза, близкому к случайному угадыванию), и при $\alpha_1 \approx 1$ (близкой к оптимальной точности прогноза), дает разное увеличение точности прогноза. Поэтому будем использовать также приближение прогноза к оптимуму $(\alpha_2 - \alpha_1)/(1 - \alpha_1)$.

Будем говорить, что точность прогноза незначительно улучшилось, или ухудшилось, если приближение к оптимуму $|(\alpha_2 - \alpha_1)/(1 - \alpha_1)| < 0.1$. При приближении к оптимуму из полуинтервала $[0.1,0.2)$, говорим, что качество прогноза умеренно увеличилось. А при приближении к оптимуму в 0.2 раза и более, говорим, что имеет место существенное увеличение точности.

6.1 Восстановление пропущенных значений и точность классификации

Сравним точность классификации при восстановлении пропущенных значений методом MICE и при обучении модели машинного обучения на некомплектной выборке. Для сравнения используем выборку breast cancer Wisconsin, из которой удалены случайным образом 40% значений. Обучим на ней ML-модель категориального бустинга CatBoost [25]. Для оценки точности прогноза здесь и далее, используем метрику сбалансированная точность, оцененную с помощью вложенной кросс-валидации.

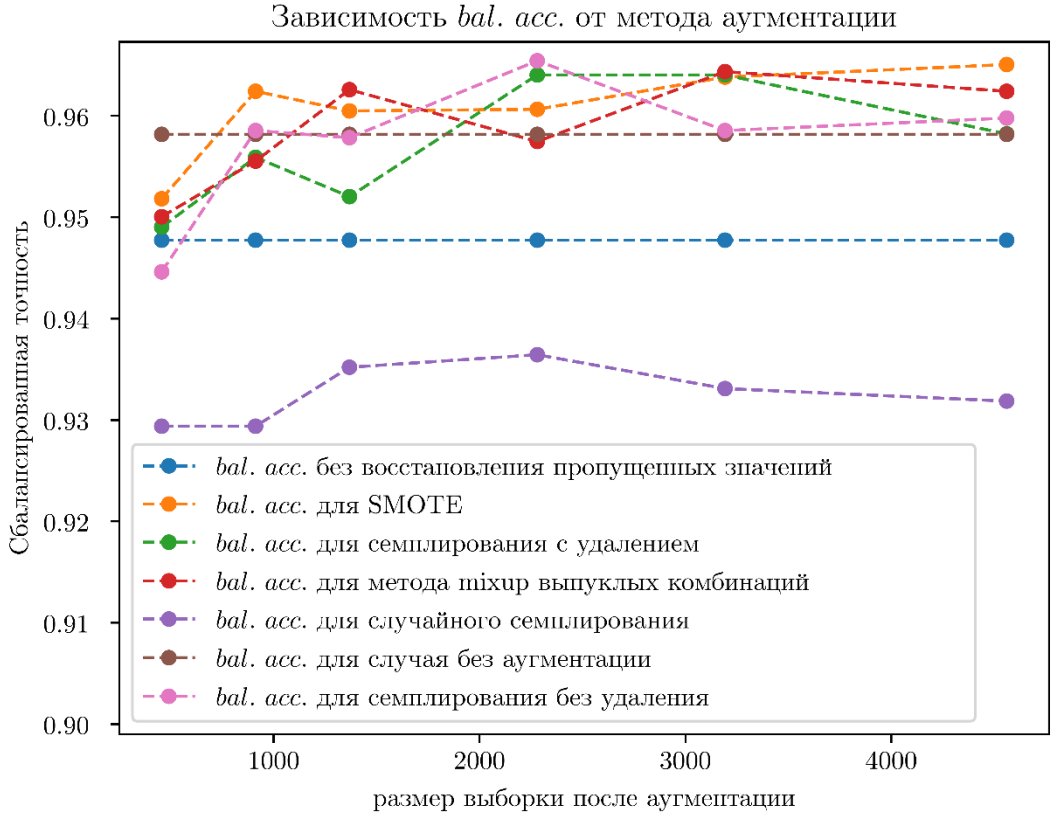

Рис. 4. Сравнение точности классификации при восстановлении пропущенных значений и при обучении ML-модели CatBoost на некомплектной выборке, а также при аугментации выборки.
Fig. 4. Accuracy comparison when learning on missing data and when using data imputation.

Сравним сбалансированную точность в случае восстановления пропущенных значений (коричневые точки) и в случае обучения на исходной некомплектной выборке (синие точки). В рамках использования схемы вложенной кросс-валидации, обучаем 20 раз на выборке, содержащей 512 объектов, строим прогноз на оставшихся 57 значениях, результаты которого сравниваем с известными ответами для получения сбалансированной точности.

- Некомплектная выборка. Сбалансированная точность: 0.948.

- Выборка, значения которой восстановлены с помощью метода MICE цепных уравнений. Сбалансированная точность: 0.959.
- Аугментация SMOTE. Сбалансированная точность изменяется от 0.96 до 0.965 в зависимости от размера выборки после аугментации.

Таким образом, при восстановлении пропущенных значений, сбалансированная точность прогноза увеличивается на 1.1%. При этом, приближаемся к оптимуму на 22%. Таким образом, восстановление пропущенных значений позволяет существенно увеличить точность прогноза. В дополнение этого, при использовании SMOTE для аугментации и балансировки выборки, получим увеличение сбалансированной точности от 0.1% до 0.6%. Это позволяет умеренно увеличить точность прогноза.

6.2 Аугментация выборки и точность классификации

Сравним влияние метода аугментации выборки на точность прогноза. В приведенной ниже таблице 2, сравним различные методы аугментации выборки между собой и со случаем отсутствия аугментации. Так как при аугментации также проводим балансировку выборки, то говорить об увеличении точности классификации из-за аугментации будем только тогда, когда после аугментации увеличивается точность как по сравнению со случаем отсутствия аугментации, так и со случаем балансировки сэмплированием без удаления.

Заметим, что предложенный метод аугментации сэмплированием с удалением позволяет увеличить сбалансированную точность и коэффициент корреляции Мэтьюса прогноза для выборок малого размера: prodromal PD и acute leukemia. При этом, обучается ML-модель опорных векторов SVM для классификации. Для выборок большего размера, наиболее эффективны методы аугментации SMOTE и mixup. При этом, обучаются ML-модели градиентного и категориального бустинга.

Для полностью заполненной выборки MNIST, аугментация данных не позволяет увеличить точность прогноза. Для несбалансированной выборки PPMI, наибольшее увеличение точности достигается из-за балансировки выборки, а не из-за ее аугментации.

Также отметим, что для выборок EyeState, breast cancer Wisconsin и cardiotocography, наибольшее относительное увеличение точности прогноза из-за аугментации достигается при обучении на некомплектной выборке. Так, для разреженной выборки EyeState, аугментация позволяет умеренно увеличить точность прогноза, а для полностью заполненной выборки, изменение точности прогноза незначительное. Для разреженной выборки Cardiotocography, аугментация позволяет существенно увеличить точность прогноза, а для полностью заполненной выборки, увеличение точности прогноза умеренное.

7. Заключение

Задача обучения на некомплектной выборке сформулирована в рамках теории ВПК-обучения. Исследована ВПК-обучаемость задачи обучения на некомплектной выборке. Был оценен рост VC-размерности выборки при восстановлении пропущенных значений фиксированным значением из конечного ограниченного множества. Доказано, что требуемый размер выборки для ВПК-обучения с фиксированной точностью, растет логарифмически медленно с ростом размера этого множества. Было доказано, что при заполнении пропущенных значений элементами сетки, требуемый размер выборки $m(\epsilon, \delta)$ для обучения с заданной точностью ϵ с вероятностью не менее $1-\delta$, увеличится незначительно по сравнению со случаем полностью заполненной выборки.

Для разреженной выборки breast cancer Wisconsin, сравнивалась точность классификации при восстановлении пропущенных значений методом MICE цепных уравнений и при обучении модели машинного обучения CatBoost на некомплектной выборке. Восстановление пропущенных значений позволило существенно увеличить точность прогноза.

Таблица 2. Сравнение методов аугментации выборки со случаем отсутствия аугментации.
Table 2. Comparison of augmentation methods with the case of no augmentation.

Название выборки	Доля пропущенных значений	ML-модель	Метод восстановления пропущенных значений	Сбалансированная точность без аугментации	Лучшая сбалансированная точность с аугментацией	Лучший метод аугментации	Приближение точности к оптимальной
Prodromal PD	23.7%	SVM	Восстановление пропущенных значений средним	0.765	0.825	сэмплирование с удалением	25.6%
Acute Leukemia	0%	SVM	Восстановление пропущенных значений средним	0.861	0.933	случайное сэмплирование	52.1%
Acute Leukemia	40%	SVM	Восстановление пропущенных значений средним	0.825	0.875	сэмплирование с удалением	28.7%
PPMI	3.8%	gradient boosting	MICE	0.661	0.667	сэмплирование без удаления	1.5%
MNIST	0%	CatBoost	Восстановление пропущенных значений средним	0.925	0.926	сэмплирование без удаления	1.3%
EyeState	0%	gradient boosting	MICE	0.922	0.924	SMOTE	2.9%
EyeState	40%	gradient boosting	MICE	0.879	0.891	SMOTE	10.1%
Breast Cancer Wisconsin	0%	gradient boosting	MICE	0.944	0.961	mixup	30.1%
Breast Cancer Wisconsin	30%	gradient boosting	MICE	0.926	0.954	SMOTE	38.2%
Cardiotocography	0%	gradient boosting	MICE	0.978	0.982	SMOTE	18.9%
Cardiotocography	40%	gradient boosting	MICE	0.944	0.966	mixup	39.5%

Аугментация и балансировка позволяют увеличить точность прогноза для всех 4 рассмотренных некомплектных средних и больших выборок, содержащих более 500 объектов. Тогда как только для 2 из 4 полностью заполненных выборок размера более 500: breast cancer Wisconsin и cardiotocography, аугментация позволяет увеличить точность классификации. Наибольшее увеличение точности аугментация позволяет достигнуть для некомплектной выборки. Отсюда можно сделать вывод о том, что аугментацию целесообразно использовать при обучении на некомплектной выборке.

Список литературы / References

[1]. Little, Roderick JA, and Donald B. Rubin. Statistical analysis with missing data. John Wiley & Sons, 2002. 389 p.

[2]. Thomas, Rajat M., et al. Dealing with missing data, small sample sizes, and heterogeneity in machine learning studies of brain disorders. Machine learning, Academic Press, 2020, pp. 249-266. DOI: 10.1016/B978-0-12-815739-8.00014-6.

[3]. Marek, Kenneth, et al. "The Parkinson progression marker initiative (PPMI). Progress in neurobiology, 95.4, 2011, pp. 629-635. DOI: 10.1016/j.pneurobio.2011.09.005.

[4]. Katunina, Elena A., et al. Searching for biomarkers in the blood of patients at risk of developing Parkinson's disease at the Prodromal Stage. International Journal of Molecular Sciences, 24.3, 2023, pp. 1842-1860. DOI: 10.3390/ijms24031842.

[5]. Golub, Todd R., et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science, 286.5439, 1999, pp. 531-537. DOI: 10.1126/science.286.5439.53.

[6]. Deng, Li. "The mnist database of handwritten digit images for machine learning research [best of the web]. IEEE signal processing magazine, 29.6, 2012, pp. 141-142. DOI: 10.1109/MSP.2012.2211477.

[7]. Campos, D. & Bernardes, J. Cardiotocography [Dataset]. UCI Machine Learning Repository, 2000. DOI: 10.24432/C51S4N.

[8]. Roesler, O. EEG Eye State [Dataset]. UCI Machine Learning Repository, 2013. DOI: 10.24432/C57G7J.

[9]. Wolberg William, H., W. N. Street, and O. L. Mangasarian. Breast cancer wisconsin (diagnostic) data set, 1995. DOI: 10.24432/C5DW2B.

[10]. Liu, Tongyu, et al. Adaptive data augmentation for supervised learning over missing data. Proceedings of the VLDB Endowment, 14.7, 2021, pp. 1202-1214.

[11]. Han, Dongmei, Qigang Liu, and Weiguo Fan. A new image classification method using CNN transfer learning and web data augmentation. Expert Systems with Applications, 95, 2018, pp. 43-56. DOI: 10.1016/j.eswa.2017.11.028.

[12]. Nanni, Loris, Gianluca Maguolo, and Michelangelo Paci. Data augmentation approaches for improving animal audio classification. Ecological Informatics, 57, 2020, pp. 1-26. DOI: 10.1016/j.ecoinf.2020.101084.

[13]. Zhou, Yue, et al. A survey on data augmentation in large model era. arXiv preprint, 2024, pp. 1-33. DOI: 10.48550/arXiv.2401.15422.

[14]. Zhang, Hongyi. Mixup: Beyond empirical risk minimization. arXiv preprint, 2017, pp. 1-13. DOI: 10.48550/arXiv.1710.09412.

[15]. Chawla, Nitesh V., et al. SMOTE: synthetic minority over-sampling technique. Journal of artificial intelligence research, 16, 2002, pp. 321-357. DOI: 10.1613/jair.953.

[16]. Valiant, Leslie G. A theory of the learnable. Communications of the ACM, 27.11, 1984, pp. 1134-1142.

[17]. Vapnik, Vladimir N., and A. Ya Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. Measures of complexity: festschrift for Alexey Chervonenkis. Cham: Springer International Publishing, 2015, pp. 11-30.

[18]. Michael, Loizos. Partial observability and learnability. Artificial Intelligence, 174.11, 2010, pp. 639-669.

[19]. Blumer, Anselm, et al. Learnability and the Vapnik-Chervonenkis dimension. Journal of the ACM (JACM), 36.4, 1989, pp. 929-965.

[20]. Yoon, Jinsung, et al. Vime: Extending the success of self-and semi-supervised learning to tabular domain. Advances in Neural Information Processing Systems, 33, 2020, pp. 1-11.

[21]. Gondara, Lovedeep, and Ke Wang. Mida: Multiple imputation using denoising autoencoders. Pacific-Asia conference on knowledge discovery and data mining. Springer International Publishing, 2018, pp. 260-272. DOI: 10.1007/978-3-319-93040-4_21.

[22]. Nelwamondo, Fulufhelo V., Shakir Mohamed, and Tshilidzi Marwala. Missing data: A comparison of neural network and expectation maximization techniques. Current Science, 2007, pp. 1514-1521. DOI: 10.48550/arXiv.0704.3474.

[23]. Rubin, Donald B. Multiple imputation. Flexible imputation of missing data, second edition. Chapman and Hall/CRC, 2018, pp. 29-62.

[24]. Campagner, Andrea. Missing but not Missed: On Learnability Under Imputation. Preprint, 2025, 1-18. DOI: 10.1007/978-3-032-06078-5_20.

[25]. Prokhorenkova, Liudmila, et al. CatBoost: unbiased boosting with categorical features. Advances in neural information processing systems, 31, 2018, pp. 1-11.

[26]. Troyanskaya, Olga, et al. Missing value estimation methods for DNA microarrays. Bioinformatics, 17.6, 2001, pp. 520-525. DOI: 10.1016/B978-0-12-815739-8.00014-6.

[27]. Li, Dan, et al. Towards missing data imputation: a study of fuzzy k-means clustering method. Springer Berlin Heidelberg, 2004, pp. 1-5. DOI: 10.1007/978-3-540-25929-9_70.

[28]. Oba, Shigeyuki, et al. A Bayesian missing value estimation method for gene expression profile data. Bioinformatics 19.16, 2003, pp. 2088-2096.

[29]. Schmitt, Peter, Jonas Mandel, and Mickael Guedj. A comparison of six methods for missing data imputation. Journal of biometrics & biostatistics, 6.1, 2015, pp. 1-7. DOI: 10.17485/ijst/2017/v10i19/110646.

[30]. Pereira, Ricardo Cardoso, Pedro Henriques Abreu, and Pedro Pereira Rodrigues. Vae-bridge: Variational autoencoder filter for bayesian ridge imputation of missing data. 2020 International Joint Conference on Neural Networks (IJCNN). IEEE, 2020. pp. 1-14.

[31]. Pereira, Ricardo Cardoso, et al. Imputation of data Missing Not at Random: Artificial generation and benchmark analysis. Expert Systems with Applications, 249, 2024, pp. 1-14. DOI: 10.1016/j.eswa.2024.123654.

[32]. Choudhury, Arkopal, and Michael R. Kosorok. Missing data imputation for classification problems. arXiv preprint, 2020, pp. 1-27. DOI: 10.48550/arXiv.2002.10709.

[33]. Stekhoven, Daniel J., and Peter Bühlmann. MissForest—non-parametric missing value imputation for mixed-type data. Bioinformatics, 28.1, 2012, pp. 112-118. DOI: 10.1093/bioinformatics/btr597.

[34]. Matthews, Brian W. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochimica et Biophysica Acta (BBA)-Protein Structure, 405.2, 1975, pp. 442-451.

[35]. Cawley, Gavin C., and Nicola LC Talbot. On over-fitting in model selection and subsequent selection bias in performance evaluation. The Journal of Machine Learning Research, 11, 2010, pp. 2079-2107.

[36]. Berrar, Daniel. Cross-validation. 2018, pp. 542-554. DOI: 10.1016/B978-0-12-809633-8.20349-X.

[37]. Stone, Mervyn. "Cross-validatory choice and assessment of statistical predictions. Journal of the royal statistical society: Series B (Methodological), 36.2. 1974, pp. 111-133. DOI: 10.1111/j.2517-6161.1974.tb00994.x.

[38]. Vabalas, Andrius, et al. Machine learning algorithm validation with a limited sample size. PloS one 14.11, 2019, pp. 1-20. DOI: 10.1371/journal.pone.0224365.

[39]. Bishop, Chris M. Training with noise is equivalent to Tikhonov regularization. Neural computation, 7.1, 1995, pp. 108-116. DOI: 10.1162/neco.1995.7.1.108.

[40]. Вапник, В.Н.., and. Червоненкис А.Ю. Теория распознавания образов: статистические проблемы обучения. Наука, 1974, 416 p. (in Russian).

[41]. Haussler, David. Decision theoretic generalizations of the PAC model for neural net and other learning applications. Information and computation, 100.1, 1992, pp. 78-150.

[42]. Sauer, Norbert. On the density of families of sets. Journal of Combinatorial Theory, Series A, 13.1, 1972, pp. 145-147. DOI: 10.1016/0097-3165(72)90019-2.

Приложение 1. Элементы теории ВПК-обучения

Приведем, следуя работе [17] Вапника и Червоненкиса, определение агностически ВПК-обучаемого множества гипотез $\mathcal{H}$. Пусть X – пространство признаков, а Y – пространство ответов. Предположим, что объекты выборки (x_i, y_i), независимы в совокупности и распределены согласно вероятностному распределению D для всех i. Пусть $l: Y \times Y \to \mathbb{R}$ – функция, с помощью которой строим функцию потерь. Наиболее распространенным вариантом l, соответствующим функции потерь $accuracy$, является (характеристическая) функция, принимающая на вход две переменных, и равная 1 в случае их равенства и 0 иначе. Определим функцию потерь L_D для гипотезы $h \in \mathcal{H}$ следующим образом: $L_D(h) = E_{(x,y)\sim D}(l(h(x), y))$. Определим функцию потерь по множеству гитотез $L_D(\mathcal{H})$ как $inf\{L_D(h): h \in \mathcal{H}\}$. Данную величину будем называть ошибкой аппроксимации множества гипотез $\mathcal{H}$ по распределению D.

Определение 1 (Агностическое ВПК обучение, Вапник и Червоненкис [17, 40]; Хаусслер [41]). Пусть X и Y – пространства признаков и ответов соответственно. Пусть $\mathcal{H}: X \to Y$ – множество гипотез. Тогда называем алгоритм A агностическим ВПК-обучателем для множества $\mathcal{H}$ тогда и только тогда, когда существует функция сложности $m: (0,1)^2 \to \mathbb{Z}^+$, такая, что для каждой пары параметров $\epsilon, \delta \in (0,1)$, для каждого распределения D над $X \times Y$,

существует $m(\epsilon,\delta) \in \mathbb{Z}^+$, такое, что для выборки $S_m = \{(x_i, y_i)_{i=1}^m\}$: $(x_i, y_i) \sim D \quad \forall 1 \le i \le m$ размера m, состоящей из m независимых в совокупности объектов x_i и ответов y_i, случайно выбранных согласно распределению D, алгоритм A, получив выборку S_m, с вероятностью не менее $1-\delta$, возвращает такую гипотезу $h \in \mathcal{H}$, для которой выполняется:

$L_D(h) \le L_D(\mathcal{H}) + \epsilon$.

Задача ВПК обучения является частным случаем агностического ВПК обучения, в котором существует оптимальная гипотеза $h \in \mathcal{H}$, для которой $h(x_i) = y_i \; \forall \, 1 \le i \le m$.

Если для множества гипотез $\mathcal{H}$ существует агностический ВПК-обучатель, то множество гипотез $\mathcal{H}$ называем агностически ВПК-обучаемым.

Определение агностической ВПК-обучаемости из работы [41] требует также полиномиальности вычислительной сложности алгоритма A от ϵ и $log\ (1/\delta)$. Однако данный аспект мы не будем рассматривать в настоящей работе.

Дадим определение размерности Вапника-Червоненкиса, или VC-размерности для класса функций, тесно связанное с ВПК-обучаемостью.

Определение 2 (Разбиение множества классом функций). Пусть $\mathcal{H}$ – класс характеристических функций $h: X \to \{0,1\}$. Класс функций $\mathcal{H}$ разбивает множество $S \subseteq X$, если для каждого подмножества $S' \subseteq S$, существует такая функция $h' \in \mathcal{H}$, которая принимает единичные значения на S' и нулевые на $S \backslash S'$.

Определение 3 (VC-размерность). Пусть $\mathcal{H}$ – класс функций $h: X \to \{0,1\}$. VC-размерность $\mathcal{H}$ – наибольший размер подмножества $S \subseteq X$, который разбивается классом $\mathcal{H}$. Обозначаем VC-размерность класса $\mathcal{H}$ как $VCdim(\mathcal{H})$.

В [42] была доказана фундаментальная теорема ВПК обучения, утверждающая, в частности, что множество гипотез $\mathcal{H}$, отображающих множество X в $\{0,1\}$, агностически ВПК-обучаемо тогда и только тогда, когда $VCdim(\mathcal{H}) < +\infty$.

Определение 4 (Функция роста). $\tau_{\mathcal{H}}: \mathbb{N} \to \mathbb{N}$ – функция роста для множества гипотез $\mathcal{H}$, содержащего характеристические функции $h: X \to \{0,1\}$, если и только если $\tau_{\mathcal{H}}(m)$ – максимальное по всем множествам $A \subseteq X$ размера m число способов, которыми множество гипотез $\mathcal{H}$ может разделить A на 2 подмножества (одно A_0^h– переводится функцией $h \in \mathcal{H}$ в 0, A_1^h переводится в 1).

В частности, при $m \le VCdim(\mathcal{H})$, функция роста составляет $\tau_{\mathcal{H}}(m) = 2^m$.

При доказательстве фундаментальной теоремы ВПК обучения, а также при доказательстве полученных в настоящей работе результатов, используется следующая лемма [42]:

Лемма Зауэра-Шелаха. Пусть $\mathcal{H}$ – множество гипотез $h: X \to \{0,1\}$. Пусть $VCdim(\mathcal{H}) < +\infty$. Тогда, выполняется следующее ограничение сверху на функцию роста $\tau_{\mathcal{H}}(m)$ для класса $\mathcal{H}$:

$\tau_{\mathcal{H}}(m) \le \sum_{i=0}^{VCdim(\mathcal{H})} \binom{m}{i} \le \left(\frac{em}{VCdim(\mathcal{H})}\right)^{VCdim(\mathcal{H})} \quad \forall m > VCdim(\mathcal{H})$.

Приложение 2. Доказательство леммы 1

Утверждение 1 леммы очевидно ввиду детерминированного заполнения пропущенных значений признаков единственным значением соответствующего признака точки X_{fill}.

Докажем утверждение 2 леммы. Рассмотрим функцию роста $\tau_{\mathcal{H}'}$ для множества $\mathcal{H}' = \mathcal{H}'(\mathcal{H}, CFill_{const})$. Пусть $\mathcal{H}_j{}' = \mathcal{H}'(\mathcal{H}, CFill_j)$ – т.е. множество гипотез $\mathcal{H}$, определенных на некомплектной выборке после ее заполнения единственным значением $x_{fill}^j \in X_{fill}$. Тогда $\mathcal{H}' = \cup_{j=1}^k \mathcal{H}_j{}'$, стало быть, выполняется соотношение $\tau_{\mathcal{H}'}(m) \le \sum_{j=1}^k \tau_{\mathcal{H}_j{}'}(m) \; \forall m \in \mathbb{N}$. Рассмотрим $m' = 2d\ log_2\,(k^{1/d} + 31)$, где $d = VCdim(\mathcal{H})$. Т.к. $m' > VCdim(\mathcal{H})$, то, по лемме Зауэра-Шелаха, выполняется, $\tau_{\mathcal{H}_j{}'}(m') \le \left(\frac{em'}{d}\right)^d$. Таким образом, $\tau_{\mathcal{H}'}(m') \le k\left(\frac{em'}{d}\right)^d$.

Докажем, что $\tau_{\mathcal{H}\prime}(m') < 2^{m\prime}$. Отсюда будет следовать требуемое утверждение о том, что $VCdim(\mathcal{H}'(\mathcal{H}, CFill_{const})) \leq m'$.

$\frac{\tau_{\mathcal{H}\prime}(m\prime)}{2^{m\prime}} \leq \frac{k(\frac{em\prime}{d})^d}{2^{m\prime}}$. Таким образом, для получения требуемой оценки сверху на $VCdim(\mathcal{H}'(\mathcal{H}, CFill_{const}))$, достаточно доказать, что

$k(\frac{em\prime}{d})^d < 2^{m\prime}$ для выбранных значений m' и d.

$k(\frac{em\prime}{d})^d < 2^{m\prime} \Leftrightarrow k^{1/d}\frac{em\prime}{d} < 2^{m\prime/d}$. Подставив $\frac{m\prime}{d} = 2\,log_2\,(k^{1/d} + 31)$, получим $k^{1/d} 2e\,log_2\,(k^{1/d} + 31) < 2^{2log_2(k^{1/d}+31)} = (k^{1/d} + 31)^2$.

Последнее неравенство верно как следствие двух неравенств:

1. $k^{1/d} < k^{1/d} + 31$
2. $2e\,log_2\,x - x < 0$ при $x \geq 32$, где за x обозначили $k^{1/d} + 31$. Для доказательство неравенства заметим, во-первых, его истинность при $x = 32$, и, во-вторых, что производная непрерывно дифференцируемой при $x > 0$ функции в левой части составляет $(2e\,log_2\,x - x)'_x = \frac{2e}{xln2} - 1 < 0$ при $x > 8 > \frac{2e}{ln2}$ □

Информация об авторах / Information about authors

Денис Олегович ЛАЗАРЕВ является специалистом кафедры теоретической информатики Института системного программирования им. В.П. Иванникова РАН. Научные интересы включают машинное обучение, вероятностный метод и алгоритмы упаковки.

Denis Olegovich LAZAREV – specialist of the Department of Theoretical Computer Science of Ivannikov Institute for System Programming of the RAS. Research interests include machine learning, probabilistic methods and packing algorithms.

Александр Владимирович ШОКУРОВ – кандидат физико-математических наук, доцент, заведующий отделом теоретической информатики Института системного программирования им. В.П. Иванникова РАН с 2019 года. Сфера научных интересов: алгебраические структуры в полях Галуа, базисы Гребнера, модулярная арифметика, нейрокомпьютерные технологии, цифровая обработка сигналов, криптографические методы защиты информации.

Alexander Vladimirovich SHOKUROV – Cand. Sci. (Phys.-Math.), Prof., Head of the Department of Theoretical Computer Science of Ivannikov Institute for System Programming of the RAS since 2019. Research interests: algebraic structures in the Galois fields, modular arithmetic, neurocomputer technologies, Grobner bases, digital signal processing, cryptographic methods for protecting information.

Станислав Александрович ФОМИН – ведущий программист. Область научных интересов: теория сложности, алгоритмы дискретной оптимизации, верификация ПО, архитектура информационных систем.

Stanislav Alexandrovich FOMIN – leading programmer. Research interests: complexity theory, program verification, discrete optimization algorithms, information systems architecture.