DOI: 10.15514/ISPRAS-2025-37(6)-22

Метод обучения персептрона на табличных данных с пропусками

А.И. Перминов, ORCID: 0000-0001-8047-0114 <perminov@ispras.ru>
А.П. Коваленко, ORCID: 0009-0007-8777-8622 <a.p.kovalenko@ispras.ru>
Д.Ю. Турдаков, ORCID: 0000-0001-8745-0984 <turdakov@ispras.ru>

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

Аннотация. Обработка пропусков в табличных данных остаётся важной задачей при построении надёжных моделей машинного обучения. В данной работе рассматривается новый подход к заполнению пропущенных значений, основанный на идее унарной классификации. Предложенный метод использует ансамбль персептронов, обучаемых отдельно для каждого класса, для оценки правдоподобия восстанавливаемых значений относительно эмпирического носителя класса. В качестве фона используется равномерное распределение на ограниченной области признакового пространства. Это позволяет интерпретировать выход модели как аппроксимацию апостериорной вероятности принадлежности объекта к классу и использовать её в процессе итеративного заполнения пропусков и обучения классификатора. Теоретически обоснована состоятельность построенной оценки. Проведены эксперименты на синтетических двумерных выборках с пропусками, распределёнными по механизму MCAR. Полученные результаты демонстрируют преимущества предложенного подхода по сравнению с классическими методами заполнения, особенно при высокой доле пропусков и сложной геометрии классов.

Ключевые слова: пропущенные данные; заполнение пропусков; унарная классификация; персептрон; машинное обучение; байесовский классификатор; оценка апостериорной вероятности; MCAR; нейросетевая регрессия.

Для цитирования: Перминов А.И., Коваленко А.П., Турдаков Д.Ю. Метод обучения персептрона на табличных данных с пропусками. Труды ИСП РАН, том 37, вып. 6, часть 2, 2025 г., стр. 93–106. DOI: 10.15514/ISPRAS–2025–37(6)–22.

Method for Training Perceptron on Tabular Data with Missing Values

A.I. Perminov, ORCID: 0000--0001-8047-0114 <perminov@ispras.ru>
A.P. Kovalenko, ORCID: 0009-0007-8777-8622 <a.p.kovalenko@ispras.ru>
D.Y. Turdakov, ORCID: 0000-0001-8745-0984 <turdakov@ispras.ru>

Ivannikov Institute for System Programming of the Russian Academy of Sciences,
25, Alexander Solzhenitsyn st., Moscow, 109004, Russia.

Abstract. Handling missing values in tabular data remains a critical challenge for building robust machine learning models. This paper presents a novel approach to imputation based on unary classification. The proposed method employs an ensemble of perceptrons trained independently for each class to estimate the likelihood of reconstructed values with respect to the empirical support of that class. A uniform distribution over a bounded region of the feature space is used as a background model, enabling the interpretation of the model's output as an approximation of the posterior probability that an object belongs to a given class. This probabilistic interpretation is then leveraged within an iterative procedure for missing value imputation and classifier training. The theoretical validity of the proposed estimator is rigorously justified. Experiments on synthetic two-dimensional datasets with missing values generated under the MCAR (Missing Completely At Random) mechanism demonstrate the superiority of the proposed method over classical imputation techniques, particularly in scenarios with high missingness rates and complex class boundaries.

Keywords: missing data imputation; unary classification; perceptron; machine learning; Bayesian classifier; posterior probability estimation; MCAR; neural network regression.

For citation: Perminov A.I., Kovalenko A.P., Turdakov D.Y. Method for training perceptron on tabular data with missing values. Trudy ISP RAN/Proc. ISP RAS, vol. 37, issue 6, part 2, 2025, pp. 93-106 (in Russian). DOI: 10.15514/ISPRAS-2025-37(6)-22.

1. Введение

Наличие пропусков в табличных данных остаётся одной из ключевых проблем при построении прикладных моделей машинного обучения. Отсутствие значений может возникать по множеству причин – от сбоев в сборе данных до неполных анкет или отказов пользователей. Игнорирование пропусков приводит к потере данных и смещению оценок, а применение стандартных подходов к заполнению зачастую не учитывает геометрию и структуру распределения.

На практике широкое распространение получили простые эвристики, такие как заполнение средним значением, модой или ближайшими соседями. Однако при высокой доле пропусков, особенно в условиях сложной формы распределения данных, такие методы могут искажать структуру выборки и снижать качество последующей модели. Эта проблема становится особенно заметной в задачах, где классы имеют запутанную или нелинейную геометрию, например, в синтетических задачах типа "двойной спирали", колец или полуколец.

В данной работе предлагается новый метод работы с пропусками, основанный на унарной классификации. Идея заключается в том, чтобы для каждого класса обучать отдельный персептрон, способный отличать реальные объекты от искусственного "фона", порождённого равномерным распределением на компакте. Такой подход позволяет интерпретировать выход модели как оценку вероятности принадлежности объекта к классу. Это даёт возможность использовать персептрон для оценки того, насколько сгенерированное значение "правдоподобно" для конкретного класса, и тем самым достовернее заполнять пропуски.

Целью данной работы является описание предложенного метода, его теоретическое обоснование, а также экспериментальное сравнение с классическими подходами на синтетических двумерных данных с пропусками, распределёнными по механизму MCAR. Особое внимание уделяется поведению метода при высокой доле пропусков и сложной структуре распределения данных.

Статья организована следующим образом. В разделе 2 формализуется задача заполнения пропусков и вводится используемая терминология. В разделе 3 описывается метод унарной классификации и его адаптация к задаче обучения с неполными данными. В разделе 4 приводится подробное описание предлагаемого алгоритма. Раздел 5 посвящён экспериментальному исследованию и анализу результатов. В заключении обсуждаются ограничения метода и возможные направления дальнейшей работы.

2. Постановка задачи

Рассматривается задача обучения классификатора по неполным данным. Пусть задано множество объектов $D = \{X, Y\} = \{(x_i, y_i)\}_{i=1}^{n},\ x_i \in R^d,\ y_i \in \{1, \dots C\}$, в котором векторы признаков x_i могут содержать пропущенные значения. Предполагается, что пропуски распределены по механизму MAR (missing at random), то есть вероятность пропуска не зависит ни от значений признаков, ни от целевой переменной.

Обозначим через M матрицу бинарных переменных, имеющую размерность входной матрицы данных, в которой $m_{ij} = 1$, если признак j в наблюдении x_i отсутствует, и $m_{ij} = 0$ в противном случае. Пусть X_0 – присутствующие части матрицы входных признаков, соответствующие элементам $m_{ij} = 0$, а X_1 – отсутствующие части, соответствующие элементам $m_{ij} = 1$. Полагаем, что выходной вектор Y известен.

Целью является построение классификатора $f: R^d \to \{1, \dots, C\}$, обладающего высокой точностью на полной версии выборки, несмотря на наличие пропусков в обучающих данных. Для решения данной задачи предлагается итеративный подход, включающий два ключевых этапа:

- Заполнение пропусков в X_1 с использованием модели, оценивающей вероятность принадлежности сгенерированных значений к каждому из классов;
- Обучение классификатора на полученной заполненной выборке.

В отличие от традиционных методов, где пропуски заполняются один раз (например, средними значениями, модой или ближайшими соседями), предлагаемый подход предполагает многократное чередование этапов генерации и обучения. После каждой эпохи обучения параметры модели обновляются, и заполнение производится заново – с учётом обновлённой вероятностной оценки правдоподобности заполненных значений. Этот итеративный процесс продолжается до сходимости или достижения заданного количества шагов.

Заполнение на каждом шаге осуществляется на основе модели унарной классификации, которая обучается отличать носитель распределения реальных наблюдений от сгенерированных из равномерного распределения. Концепция унарной классификации и её адаптация к задаче восстановления пропусков подробно рассматриваются в следующем разделе.

3. Метод унарной классификации (случай одного класса)

Метод обучения классификатора при наличии пропущенных входных данных обучающей выборки основан на построении байесовского унарного классификатора.

В работе [1] предложен метод экстраполяции байесовского бинарного классификатора, когда к данным, относящимся к двум разным классам (с метками "+1" и "-1"), добавляется третий, искусственно созданный «фоновый» класс с меткой "0", представляющий собой случайную выборку из заданного на компакте равномерного распределения. Модифицированный таким образом байесовский классификатор, помимо решений об отнесении наблюдений по значению дискриминантной функции к первому или второму классу, может принимать решение об отказе от классификации при близости значений дискриминантной функции к

нулю. Доказано, что на носителе распределения модифицированный байесовский классификатор эквивалентен байесовскому классификатору, за пределами носителя принимается решение об отказе от классификации. Поэтому входным наблюдениям, принадлежащим компакту, но лежащим за пределами эмпирический границы носителя исходного распределения (например, выбросам), будет «отказано» в классификации.

3.1 Унарный байесовский классификатор

Унарный байесовский классификатор отличается от рассмотренного выше модифицированного бинарного байесовского классификатора тем, что во входной выборке присутствуют наблюдения только одного класса, которые обозначены меткой «1», а наблюдения из второго, «фонового» класса обозначены меткой «0». Формально задача унарной классификации может быть представлена следующим образом.

Предположим, что входные данные представляют собой наблюдения n независимых одинаково распределенных случайных d-мерных векторов из R^d, имеющих равномерно непрерывную плотность распределения $f(x)$. Предположим также, что носитель распределения $f(x)$ неизвестен, но расположен внутри компакта $K = [0,1]^d$.

Пусть задана случайная величина (X, Y), где X – d-мерный случайный вектор с равномерно непрерывной плотностью смеси распределений $\alpha f(x) + (1-\alpha)p(x)$, где $f(x)$ – плотность распределения наблюдений исходной выборки, $p(x)$ – плотность равномерного распределения наблюдений «фона», заданная на компакте K, α – весовой коэффициент, $0 \leq \alpha \leq 1$, метка Y принимает значение 1 или 0 в зависимости от того, какому классу принадлежит вектор X, то есть выборке или «фону».

Пусть

$$g(x) = P(Y = 1 \vee X = x) = E(Y \vee X = x) = \frac{f(X)}{f(X) + \frac{1-\alpha}{\alpha} \cdot p(X)} \quad (1)$$

есть апостериорная вероятность выборочного класса (функция регрессии Y на X). Если функция $g(x)$ известна, то задача унарной классификации может быть решена следующим образом: если в точке x функция $g(x) > 0$, то эта точка принадлежит носителю распределения, то есть выборочному классу, в противном случае – это «выброс», относительно которого решение принимается некоторой дополнительной процедурой. Однако, функция $g(x)$, как правило, неизвестна и требуется построить ее аппроксимацию по имеющейся выборке данных.

3.2 Построение аппроксимации

Пусть $c(x)$ – непрерывная функция, заданная на компакте K.

Рассмотрим задачу среднеквадратической аппроксимации (минимизация осуществляется по всем $c(X)$):

$$c^*(x) = argmin\, E(c(x) - Y)^2 \quad (2)$$

Поскольку $E(c(x) - Y)^2 = E(c(x) - g(x) + g(x) - Y)^2 = E\big(c(x) - g(x)\big)^2 + E(g(x) - Y)^2$, а второе слагаемое от $c(x)$ не зависит, задача (2) эквивалентна аппроксимации равномерно непрерывной функции регрессии Y на X:

$$c^*(x) = argmin\, E\big(c(x) - g(x)\big)^2 \quad (3)$$

3.3 Аппроксимация с помощью персептрона

В качестве функции $c(x) = c(x; k, L)$ рассмотрим многослойный персептрон (полносвязную нейросеть) с кусочно-линейной функцией активации $|\cdot|$, состоящий из L скрытых слоев по

k нейронов в каждом. По основной аппроксимационной теореме [2] для любого заданного $\epsilon > 0$ существуют такие значения параметров персептрона k и L, что для любого $x \in K$ выполняется условие (4):

$$\sup_{x \in K} |c(x) - g(x)| < \epsilon \tag{4}$$

то есть теоретически ϵ-приближенное решение задачи (3) существует.

Рассмотрим выборочную постановку задачи унарной классификации. Пусть задана выборка $\{(X_1, Y_1), (X_2, Y_2), \cdots, (X_n, Y_n)\}$, где $X_i \in K, Y_i = 1$, которую будем интерпретировать как размеченный набор n наблюдений случайных векторов, распределенных с равномерно непрерывной плотностью $f(x)$. Для формирования выборки из смеси выборочного класса и «фона» с плотностью $\alpha f(x) + (1-\alpha)p(x)$ добавим к этой выборке искусственно сгенерированные данные $\{(X_{n+1}, Y_{n+1}), (X_{n+2}, Y_{n+2}), \cdots, (X_{n+m}, Y_{n+m})\}$, где $m = n \cdot \frac{1-\alpha}{\alpha}$, векторы $X_{n+i}, i = 1,2, \cdots m$, есть наблюдения независимых равномерно распределённых на компакте K случайных векторов, $Y_{n+i} = 0$.

Пусть $C(k, L)$ – множество всех многослойных персептронов $c(x)$ с кусочно-линейной функцией активации $|\cdot|$ в скрытых слоях и числом L и размером k скрытых слоев.

Применяя некоторый алгоритм оптимизации, построим выборочную оценку решения задачи (2):

$$\sum_{i=1}^{n+M} (c_n(X_i) - Y_i)^2 \to min \tag{5}$$

где минимизация функционала осуществляется по всем $c_n(X) \in C(l, L)$, а параметры k и L выбраны оптимально с учетом ограничений, связанных с переобучением.

Пусть функция $c_n^*(X)$ есть решение оптимизационной задачи (5), которую будем называть **функцией нейросетевой регрессии**. Соответствующий этому решению персептрон строит иерархическое разбиение компакта K на N непересекающихся ячеек $K = \{K_1, K_2, \cdots, K_N\}$ [3] (рис. 1).

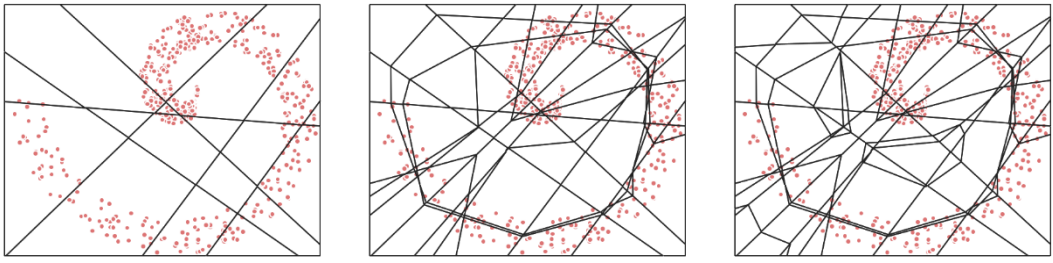

Рис. 1. Пример разбиения некоторым MLP с $L = 2$, $k = 6$.
Fig. 1. Some MLP partition example, $L = 2$, $k = 6$.

3.4 Состоятельность метода унарной классификации

Для обоснования состоятельности решения $c_n^*(X)$ рассмотрим кусочно-постоянную (в общем случае разрывную) функцию гистограммной регрессии $h_n(X)$, и решим оптимизационную задачу:

$$\sum_{i=1}^{n+M} (h_n(X_i) - Y_i)^2 \to min \tag{6}$$

где минимизация осуществляется по всем кусочно-постоянным функциям, принимающим постоянные значения в ячейках разбиения компакта $K = \{K_1, K_2, \cdots, K_N\}$.

Пусть $X \in K_r$. Тогда задачу (6) для этой ячейки можно представить в виде:

$$n_1(X) \cdot (h_n(X) - 1)^2 + n_0(X) \cdot (h_n(X) - 0)^2 \to min, \tag{7}$$

где $n_1(X) = \sum_{i=1}^{n+m} I_{X_i \in K_r, Y_i = 1}$, $n_0(X) = \sum_{i=1}^{n+m} I_{X_i \in K_r, Y_i = 0}$.

После дифференцирования функции (7) по $h_n(X)$ получаем решение задачи (6):

$$h_n^*(X) = \frac{n_1(X)}{n_1(X) + n_0(X)} = \frac{f_n(X)}{f_n(X) + \frac{1-\alpha}{\alpha} \cdot p_n(X)}, \tag{8}$$

где $f_n(X) = \frac{n_1(X)}{n \cdot \mu(K_r)}$ – адаптивная гистограммная оценка плотности $f(x)$ в ячейке K_r, $p_n(X) = \frac{n_0(X)}{n \cdot \mu(K_r)}$ – адаптивная гистограммная оценка равномерной плотности в ячейке K_r, $\mu(K_r)$ – мера ячейки K_r. Пример вычисления функции гистограммной регрессии показан на рис. 2.

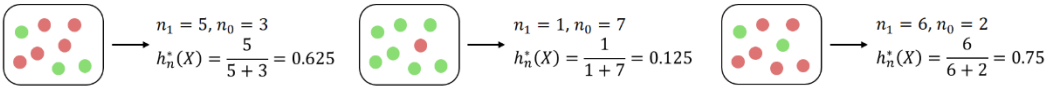

Рис. 2. Пример вычисления $h_n^*(X)$ в некоторой ячейке K_r.
Fig. 2. $h_n^*(X)$ evaluation example at some cell K_r.

В работах [4-5] сформулированы асимптотические условия строгой состоятельности адаптивных гистограммных оценок плотности распределения, при выполнении которых для любого произвольно малого $\epsilon > 0$ и $X \in K$ с вероятностью 1 имеют место соотношения:

$$\begin{aligned} |f(X) - f_n(X)| < \epsilon \\ |p(X) - p_n(X)| < \epsilon \end{aligned} \tag{9}$$

Отсюда следует, что при достаточно больших n для любого $X \in K$ с вероятностью 1:

$$|g(X) - h_n^*(X)| < \epsilon_1, \tag{10}$$

где $\epsilon_1 = \frac{4\epsilon}{\mu(K)}$, $\mu(K)$ – мера компакта K.

Для состоятельности, в частности, требуется, чтобы диаметр ячеек убывал с ростом n, но при этом число точек внутри ячеек стремилось к бесконечности. При размерности пространства $d = 10$ эти требования выполняются уже при малых значениях k и L. Например, для нормальной плотности распределения, $d = 10$, $k = 10$ и $L = 2$ количество ячеек N превышает десятки тысяч, а для их 90%-го заполнения (чтобы в ячейку попала хотя бы одна фоновая точка) требуются миллионы фоновых точек. Следует отметить, что в «заполненных» ячейках, в которых представлены как выборочные, так и фоновые точки, значения функций $h_n^*(X)$ и $c_n^*(X)$ близки (рис. 3). В ячейках, в которые попали только фоновые точки, гистограммная регрессия $h_n^*(X)$, а значение нейросетевой регрессии $c_n^*(X)$ определяется интерполяцией значений в соседних ячейках благодаря непрерывности функции. В областях высокой плотности $f(X)$ значения $c_n^*(X)$ существенно больше нуля, в областях низкой плотности нейросетевая регрессия близка к нулю.

Поэтому при достаточно больших n, правильно подобранных значениях параметров сети k и L и, учитывая соотношения (3) - (10), нейросетевую регрессию $c_n^*(X)$ можно считать состоятельной оценкой апостериорной вероятности $g(X)$, а значит, при решении прикладных задач есть основания полагать, что имеет место соотношение:

$$c_n^*(X) \approx h_n^*(X) \approx g(X) = P(Y = 1 \vee X), \tag{11}$$

то есть значение нейросетевой регрессии можно рассматривать как оценку апостериорной вероятности выборочного класса.

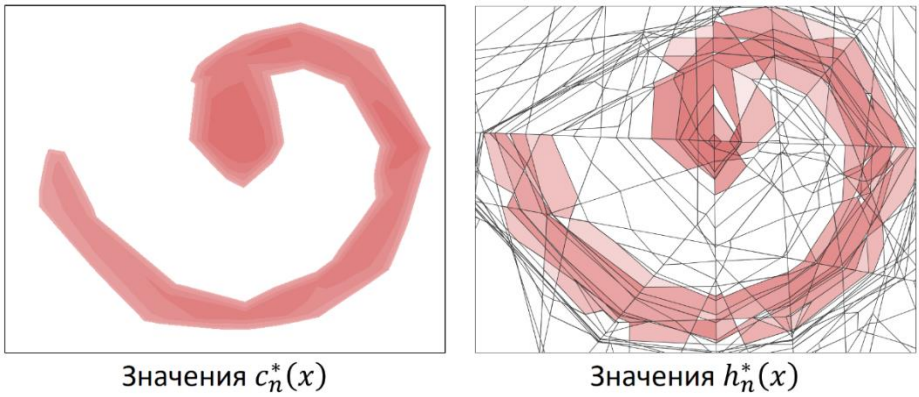

Рис. 3. Сравнение $c_n^*(X)$ и $h_n^*(X)$.
Fig. 3. Comparison of $c_n^*(X)$ and $h_n^*(X)$.

4. Метод обучения MLP при наличии пропусков в обучающей выборке.

Представим входные данные в виде объединения выборок, соответствующим отдельным классам (то есть в каждой такой выборке метки наблюдений совпадают). Пусть число классов равно C. Каждую выборку, в свою очередь, разделим на две подвыборки: первая состоит из наблюдений без пропусков признаков (комплектная подвыборка класса), а вторая – из наблюдений с пропусками (некомплектная подвыборка класса). Обучение MLP осуществляется в соответствии с предположением MAR независимо в каждом классе.

Предлагаемый метод обучения MLP при наличии пропусков в обучающей выборке применяется последовательно к каждому из C классов и состоит из трёх шагов (схематичная иллюстрация к методике дана на рис. 4).

1) **Начальное обучение.** Для комплектной подвыборки $\{X_i, i = 1,2,\cdots,n\}$ j-го класса, $j \in \{1,2,\cdots,C\}$, решить задачу унарной классификации и построить MLP_j, реализующий кусочно-линейную непрерывную функцию $c_n^j(X)$, заданную на компакте K.
2) **Дообучение.** Дообучение осуществляется по всей обучающей выборке j-го класса отдельными эпохами. Перед текущей эпохой выполнить временное (для данной эпохи) заполнение некомплектных наблюдений. Для каждого некомплектного наблюдения X:
   a. Разделить множество индексов координат вектора $X = (x_1, x_2, \cdots, x_d)$ на два подмножества M_0 и M_1, включающие соответственно индексы заполненных и пропущенных координат.
   b. Заполнить координаты X из M_1 наблюдениями равномерно распределенной случайной величины на отрезке $[0,1]$, в результате чего будет получен комплектный вектор X'. Вычислить $c_n^j(X')$. Сгенерировать наблюдение биномиальной случайной величины с вероятностью успеха $p = c_n^j(X')$.
   c. При успешном исходе временно заменить в обучающей выборке некомплектный вектор X на комплектный вектор X' и перейти к рассмотрению следующего некомплектного наблюдения. В противном случае повторить шаг 2.b.
   d. Выполнить дообучение сети по «доукомплектованной» обучающей выборке.

3) Перейти к следующей эпохе дообучения, повторяя шаги a-d, до полного завершения обучения MLP_j для j-го класса с функцией нейросетевой регрессии $c_n^j(X)$.

Повторяя шаги 1-3 для всех классов, получим C обученных нейросетей MLP_j и соответствующих им непрерывных кусочно-линейных функций $\{c_n^1(X), c_n^2(X), \cdots, c_n^C(X)\}$, каждая из которых есть выборочная оценка апостериорной вероятности соответствующего класса в точке X.

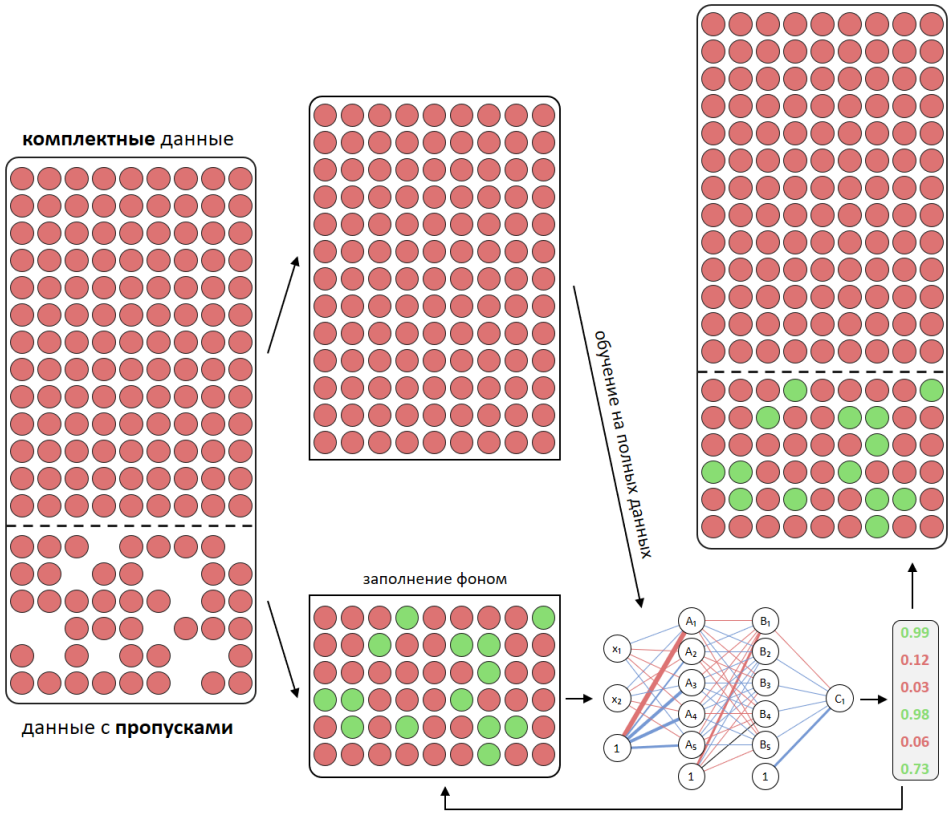

Рис. 4. Схема обучения $c_n^j(X)$ по неполным данным.
Fig. 4. $c_n^j(X)$ training on missed data diagram.

Для решения задачи классификации комплектного наблюдения X возможны различные стратегии. Простейшая состоит в выборе класса, для которого апостериорная вероятность максимальна. Другой вариант – выбрать в качестве решения все классы, значения апостериорной вероятности для которых больше некоторого заданного порога, и продолжить решение задачи классификации, например, в другом признаковом пространстве. Возможно, целесообразно учитывать априорные вероятности классов, различные функции стоимости ошибок для разных классов и т.п. Рассмотрение этих вопросов выходит за рамки данного исследования и для простоты используется выбор класса с максимальным значением $c_n^j(X)$.

Реализация описанного метода обучения и сценарии всех экспериментов доступны в открытом репозитории [6], включающим код генерации синтетических данных, обучение MLP с применением различных стратегий заполнения, а также визуализацию апостериорных распределений.

Таким образом, предложенный метод позволяет отказаться от прямого восстановления недостающих признаков, заменяя его вероятностной процедурой включения неполных

наблюдений в процессе обучения. Это особенно важно в случаях, когда форма распределения классов не допускает корректного заполнения с помощью глобальных статистик.

5. Эксперименты

Цель данного раздела – эмпирически оценить эффективность предлагаемого метода на синтетических наборах данных с различной топологической и геометрической структурой при различной доле пропусков. Особое внимание уделяется случаям, когда стандартные методы заполнения демонстрируют снижение качества из-за неспособности учесть сложную форму распределения классов.

5.1 Наборы данных

В экспериментах использовались следующие двумерные синтетические наборы данных:

- **Гауссианы** – два нормально распределённых кластера с равной дисперсией и небольшим перекрытием (рис. 5, слева);
- **Спирали** – классы формируют витки спиралей с общей точкой начала координат, разделение классов сильно нелинейное (рис. 5, по центру);
- **Кольцо и круг** – один класс расположен внутри круга, второй образует кольцо с зазором между границами (рис. 5, справа).

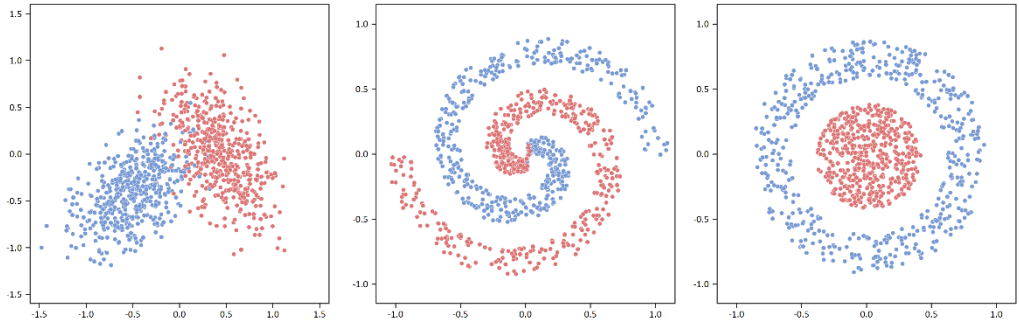

Рис. 5. Наборы данных: гауссианы (слева), спираль (центр), кольцо и круг (справа).
Fig. 5. Datasets: gaussians (left), spiral (center), ring and circle (right).

Каждый набор содержал 1000 наблюдений для обучающей части и 5000 наблюдений для тестовой.

5.2 Обработка пропущенных значений

Во всех наборах данных искусственно вводились пропуски в признаках с уровнями 20%, 40%, 50%, 60%, 80% и 90%. Пропуски вносились случайно и только в признаках (целевые метки всегда сохранялись). Были рассмотрены следующие методы обработки пропусков:

- $mean$ – заполнение по среднему значению признака.
- $mode$ – заполнение наиболее частым значением.
- kNN $(k = 3)$ – заполнение по ближайшим трём соседям в евклидовом пространстве.
- kNN $(k = 7)$ – аналогично, но с $k = 7$.
- $reproduction$ (предлагаемый метод) – метод, основанный на унарной классификации из раздела 4.

5.3 Сценарии обучения

Для каждой комбинации набора данных и уровня пропусков модель обучалась в следующих режимах:

- $full$ – обучение на полном наборе без пропусков.
- $complete$ – обучение только на тех примерах, где отсутствуют пропуски.
- $imputed$ – обучение на наборе, где пропуски заполнялись одним из методов.

В качестве модели использовался полносвязный персептрон с $L = 2$ скрытыми слоями по $k = 20$ нейронов в каждом и одним выходным слоем. Обучение осуществлялось на протяжении 500 эпох. Для метода репродукции персептрон обучался в течение 50 эпох на данных без пропусков, а затем каждую эпоху запускался процесс вероятностного заполнения пропусков и обучение продолжалось уже на обновлённых заполненных данных.

5.4 Оценка качества

Каждая комбинация набора данных, уровня пропусков и метода заполнения запускалась 50 раз с различными начальными инициализациями весовых коэффициентов. В качестве основной метрики использовалась правильность классификации (accuracy) на тестовом множестве из соответствующего набора данных из 5000 элементов. Все тестовые наборы содержали только полные данные.

5.5 Результаты

Результаты со значениями accuracy (среднее ± стандартное отклонение) по 50 запускам представлены в табл. 1, табл. 2 и табл. 3. Визуальный анализ показывает, что предлагаемый метод репродукции демонстрирует более высокую устойчивость при высоких уровнях пропусков, особенно на сложных наборах данных, как например "кольцо и круг". Традиционные методы заполнения (среднее, мода) показывают ожидаемое снижение качества, особенно при пропусках выше 60%. Метод kNN даёт умеренное улучшение, но чувствителен к плотности выборки.

Табл. 1. Результаты на наборе данных «Гауссианы».
Table 1. Results on the "Gaussians" dataset.

пропуски	full	complete	reproduce	mean	most frequent	knn 3	knn 7
20%	0.925 ± 0.003	0.919±0.006	**0.925±0.004**	0.917±0.007	0.931±0.007	0.919±0.004	0.921±0.006
40%		**0.919±0.006**	0.917±0.009	0.897±0.008	0.913±0.020	0.909±0.010	0.915±0.005
50%		0.917±0.005	**0.918±0.008**	0.904±0.014	0.909±0.018	0.906±0.007	0.917±0.005
60%		0.910±0.013	**0.921±0.009**	0.866±0.018	0.869±0.047	0.888±0.011	0.894±0.012
80%		0.875±0.011	**0.912±0.008**	0.752±0.047	0.781±0.057	0.852±0.015	0.851±0.011
90%		0.842±0.023	**0.906±0.007**	0.593±0.092	0.627±0.128	0.723±0.030	0.806±0.015

Табл. 2. Результаты на наборе данных «Спираль».
Table 2. Results on the "Spiral" dataset.

пропуски	full	complete	reproduce	mean	most frequent	knn 3	knn 7
20%	0.941 ± 0.024	0.936±0.031	**0.945±0.025**	0.922±0.032	0.934±0.029	0.927±0.034	0.941±0.020
40%		0.924±0.023	**0.930±0.021**	0.899±0.034	0.892±0.057	0.880±0.057	0.918±0.025
50%		**0.926±0.016**	0.910±0.034	0.893±0.031	0.873±0.047	0.867±0.032	0.868±0.062
60%		**0.913±0.030**	0.898±0.047	0.890±0.027	0.841±0.081	0.823±0.044	0.853±0.040
80%		**0.869±0.039**	0.861±0.044	0.750±0.095	0.632±0.127	0.727±0.041	0.773±0.062
90%		**0.827±0.042**	0.812±0.067	0.545±0.082	0.373±0.129	0.648±0.049	0.695±0.040

Табл. 3. Результаты на наборе данных «Кольцо и круг».
Table 3. Results on the "Ring and circle" dataset.

пропуски	full	complete	reproduce	mean	most frequent	knn 3	knn 7
20%	0.981 ± 0.014	0.987±0.009	**0.989±0.006**	0.941±0.058	0.983±0.008	0.954±0.042	0.927±0.086
40%		0.984±0.011	**0.986±0.011**	0.865±0.103	0.970±0.015	0.887±0.080	0.846±0.123
50%		0.971±0.019	**0.984±0.016**	0.852±0.094	0.958±0.025	0.868±0.119	0.751±0.149
60%		0.974±0.018	**0.981±0.016**	0.763±0.083	0.907±0.068	0.797±0.059	0.705±0.103
80%		0.892±0.136	**0.964±0.031**	0.609±0.148	0.673±0.177	0.781±0.058	0.618±0.069
90%		0.852±0.080	**0.952±0.038**	0.295±0.126	0.433±0.170	0.522±0.039	0.562±0.055

Для дополнительного визуального анализа на рис. 6, рис. 7 и рис. 8 показаны некоторые (лучшие) модели, полученные после обучения с помощью метода репродукции.

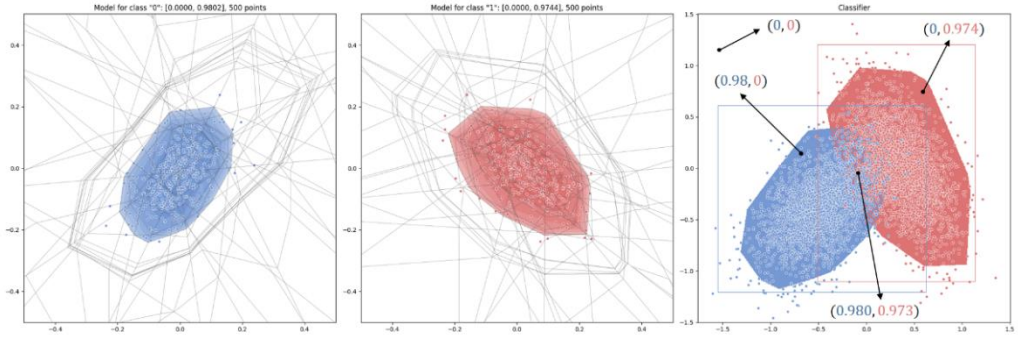

Рис. 6. Модель, полученная при классификации набора данных "гауссианы".
Fig. 6. Model obtained by classifying the gaussian dataset.

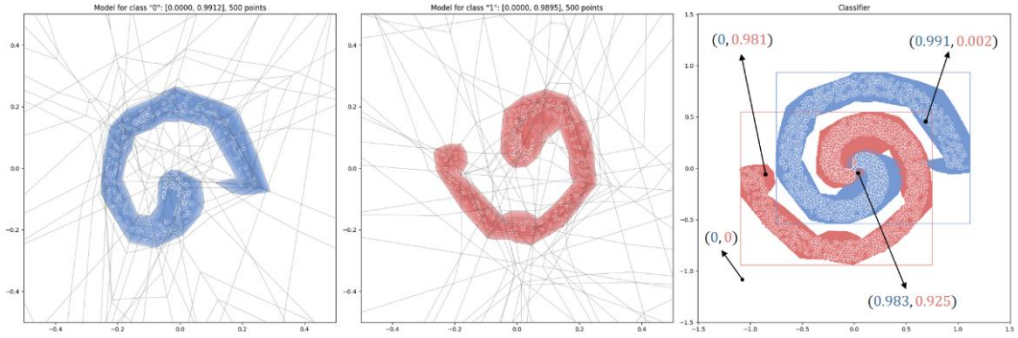

Рис. 7. Модель, полученная при классификации набора данных "спираль".
Fig. 7. Model obtained by classifying the spiral dataset.

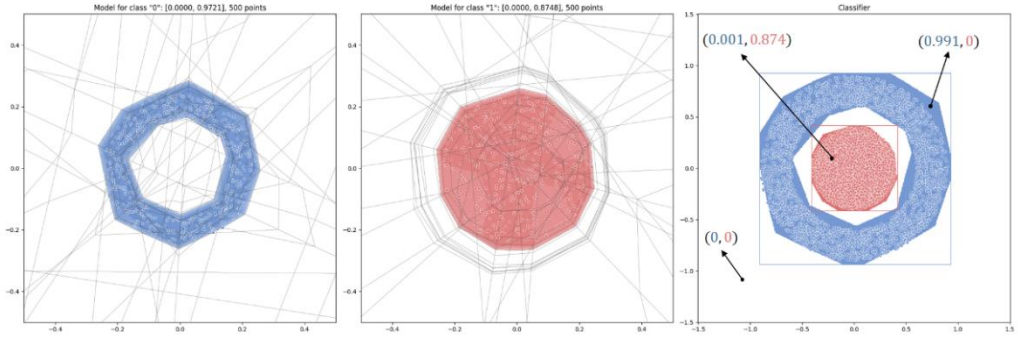

Рис. 8. Модель, полученная при классификации набора данных "кольцо и круг".
Fig. 8. Model obtained by classifying the ring and circle dataset.

Таким образом, предложенный подход воспроизведения недостающих признаков на основе унарной классификации демонстрирует высокую устойчивость к пропускам и способность к адаптации к сложной геометрии данных. Особенно заметно его преимущество на структурах, где границы классов являются нелинейными и неоднородными, как, например, в задачах «кольцо и круг» или «спирали». Метод воспроизведения не требует оценки плотности или предположений о форме распределения и теоретически может быть расширен на более высокие размерности. В сочетании с доступной реализацией и возможностью встраивания в существующие модели, он представляет собой практический и теоретически обоснованный инструмент для работы с неполными данными.

6. Заключение

В данной работе был предложен метод обработки пропусков в табличных данных, основанный на идее унарной классификации и вероятностного восстановления значений на фоне равномерного распределения. Подход ориентирован на работу с пропущенными признаками при обучении многослойного перцептрона (MLP) и направлен на корректное представление неопределённости, связанной с отсутствием информации.

Эксперименты на синтетических двумерных задачах показали, что метод воспроизведения обеспечивает более высокое качество классификации по сравнению с классическими подходами, такими как удаление объектов с пропусками, заполнение средним и kNN-интерполяция. Особенно заметным это преимущество становится в задачах с топологически сложными границами классов, где искажение геометрии данных приводит к существенной деградации качества у традиционных методов.

Предложенный метод не требует предварительного заполнения и интегрируется непосредственно в процесс обучения. Он сохраняет гибкость нейросетевых моделей, одновременно учитывая вероятностную природу отсутствующих значений.

Тем не менее, работа имеет ряд ограничений. В настоящей формулировке метод реализован для пропусков в одном признаке и на фиксированной архитектуре MLP. Кроме того, эффективность метода не проверялась на реальных табличных наборах данных с пропущенными данными, где могут действовать более сложные механизмы отсутствия информации (MAR, MNAR).

Будущая работа может быть направлена на:

- расширение метода на произвольное количество пропущенных признаков;
- адаптацию к другим типам моделей (например, градиентный бустинг, трансформеры);
- применение к реальным задачам, включая медицинские и социологические данные;
- исследование устойчивости при различных механизмах пропусков.

Таким образом, предложенный подход открывает перспективное направление для построения устойчивых моделей, способных эффективно обучаться в условиях неполных данных.

Список литературы / References

[1]. Lukyanov K. S. et al. Extrapolation of the Bayesian classifier with an unknown support of the two-class mixture distribution //Russian Mathematical Surveys. 2024. Vol. 79, No. 6, pp. 991-1015.

[2]. Cybenko G. Approximation by superpositions of a sigmoidal function //Mathematics of control, signals and systems. 1989. Vol. 2, No. 4, pp. 303-314.

[3]. Kovalenko A. Geometric interpretation of a multilayer perceptron with piecewise linear activation functions // 31st scientific and technical conference "MiTSOBIT". Saint Petersburg, 2022. pp. 34—35.

[4]. Devroye L. Nonparametric density estimation //The L_1 View. 1985.

[5]. Devroye L., Györfi L., Lugosi G. A probabilistic theory of pattern recognition. – Springer Science & Business Media, 2013. Vol. 31.
[6]. MissingDataPerceptron, https://github.com/dronperminov/MissingDataPerceptron, last accessed: 01 July 2025.

Информация об авторах / Information about authors

Андрей Игоревич ПЕРМИНОВ является аспирантом института системного программирования РАН. Научные интересы: нейросетевая обработка данных, цифровая обработка изображений, методы доверенного искусственного интеллекта.

Andrey Igorevich PERMINOV – a postgraduate student at the Institute of System Programming of the RAS. Research interests: neural network data processing, digital image processing, trusted artificial intelligence.

Андрей Петрович КОВАЛЕНКО – доктор технических наук, исследователь центра доверенного искусственного интеллекта в институте системного программирования РАН. Научные интересы: методы доверенного искусственного интеллекта.

Andrey Petrovich KOVALENKO – Dr. Sci. (Tech.), a researcher at the Center for Trusted Artificial Intelligence at the Institute of System Programming of the RAS. Research interests: trusted artificial intelligence.

Денис Юрьевич ТУРДАКОВ – кандидат физико-математических наук, заведующий отделом ИСП РАН. Научные интересы: анализ социальных сетей, анализ текста, извлечение информации, обработка больших данных, методы доверенного искусственного интеллекта.

Denis Yurievich TURDAKOV – Cand. Sci. (Phys.-Math.), head of department ISP RAS. Research interests: social network analysis, text mining, information extraction, big data, trusted artificial intelligence.