ИПМ им.М.В.Келдыша РАН • Электронная библиотека

Препринты ИПМ • Препринт № 53 за 2024 г.

ISSN 2071-2898 (Print)
ISSN 2071-2901 (Online)

Ю.С. Чайников, В.А. Судаков

Об оценках интегрального риска предиктора липшицевых функций в моделях машинного обучения

Рекомендуемая форма библиографической ссылки: Чайников Ю.С., Судаков В.А. Об оценках интегрального риска предиктора липшицевых функций в моделях машинного обучения // Препринты ИПМ им. М.В.Келдыша. 2024. № 53. 12 с. https://doi.org/10.20948/prepr-2024-53
https://library.keldysh.ru/preprint.asp?id=2024-53

О р д е н а Л е н и н а
ИНСТИТУТ ПРИКЛАДНОЙ МАТЕМАТИКИ
имени М.В. Келдыша
Р о с с и й с к о й а к а д е м и и н а у к

Ю.С. Чайников, В.А. Судаков

Об оценках интегрального риска предиктора липшицевых функций в моделях машинного обучения

Москва — 2024

Чайников Ю.С., Судаков В.А.

Об оценках интегрального риска предиктора липшицевых функций в моделях машинного обучения

Несбалансированность классов в доступных обучающих выборках при решении задач машинного обучения в большинстве практических случаев затрудняет тренировку предикторов, эффективно обобщающих закономерности обучающего датасета на генеральную совокупность. В работе исследованы теоретические основания эффективности добавления в обучающую выборку синтетических данных. В оценке общего риска выделено два типа ошибок: ошибка репрезентации и ошибка уклонения. Сформулированы практические рекомендации по созданию синтетических выборок, уклоняющихся в своем распределении от репрезентативных по плотности распределения аргумента, с более частыми сэмплами в тех областях, где плотность распределения аргумента имеет относительно малые значения, что ведет к уменьшению размера соответствующих ячеек Вороного и уменьшению вклада ошибки уклонения в общий риск.

Ключевые слова: синтетические данные, машинное обучение, ячейки Вороного, предиктор, обучающая выборка, общий риск, эмпирический риск, ошибка репрезентации, ошибка уклонения

Yuri Sergeevich Chaynikov, Sudakov Vladimir Anatolyevich

On the estimation of integral risk of predictor Lipschitz functions in machine learning models

Class imbalance in available training samples for solving machine learning problems in most practical cases complicates the training of predictors that effectively generalize patterns from the training dataset to the general population. This paper investigates the theoretical foundations of the effectiveness of adding synthetic data to the training set. In the assessment of overall risk, two types of errors are highlighted: representation error and deviation error. Practical recommendations are formulated for creating synthetic samples that deviate in their distribution from the representative ones by the density distribution of the argument, with more frequent samples in those areas where the density distribution of the argument has relatively low values, leading to a reduction in the size of the corresponding Voronoi cells and a reduction in the contribution of deviation error to total risk.

Key words: synthetic data, machine learning, Voronoi cells, predictor, training sample, total risk, empirical risk, representation error, deviation error

Введение

В работе рассматривается одна из постановок регрессионной задачи. В стандартной постановке задачи строится аппроксиматор (предиктор) f, достаточно приближенный к неизвестной функции $\hat{f}$, по известным значениям функции $\hat{f}$ на обучающих сэмплах. При этом аппроксиматор ищется из класса допустимых аппроксиматоров.

В классической постановке регрессионной задачи размер обучающей выборки зафиксирован при постановке задачи и не подлежит изменению. В широком классе практических задач улучшение качества предиктора достигается за счет увеличения размера обучающей выборки путем аугментации данных, в том числе через добавление в обучающую выборку синтетических данных.

Особенно остро проблема недостаточности данных стоит при использовании в качестве предиктора глубоких нейронных сетей. Среди подходов к аугментации данных актуальны смешанный (mixup) подход [1], использование генеративных состязательных сетей для создания дополнительных сэмплов [2], перебалансировка классов и вовлечение неразмеченных данных из той же области определения, что и размеченная выборка [3], методы аугментации на основе линейных преобразований и нелинейных искажений [4].

Обобщение сэмплированных данных от точечных дискретных значений к распределениям в небольших окрестностях исходных значений как в пространстве сэмплов, так и в латентных пространствах при использовании генеративно-состязательных сетей с последующим использованием полученных распределений для генерации новых сэмплов в обучающую выборку [5] с дополнительной валидацией [6; 7; 8; 9; 10; 11] также представляет интерес в контексте аугментации данных.

Основная часть работ в этой области направлена на изучение сугубо практических приемов аугментации данных, уделяя меньше внимания теоретическим предпосылкам эффективности этих приемов. В нашей работе мы сосредоточимся на исследовании теоретических оснований эффективности добавления в обучающую выборку синтетических данных, оставляя за рамками статьи сами методы создания синтетических данных, существенно зависящие от домена.

Постановка задачи построения аппроксиматора

Пусть имеется:

1. Множество возможных измерений $x \in X \subset \mathbb{R}^n$. Мы ограничимся рассмотрением случая, когда X – компакт, а функция плотности распределения $p(x)$ непрерывна на X.

2. «Истинная», неизвестная функция измерения $\hat{f}(x): X \to \mathbb{R}^m$. В общем случае образ $\hat{f}(x)$ лежит в $\mathbb{R}^m$ при $m \geq 1$. Однако многомерный случай отличается от случая $m = 1$ не принципиально. Рассмотрение случая $m > 1$ отличается лишь дополнительными выкладками. Далее мы считаем, что $m = 1$.
3. Обучающая выборка $S = \{x_i, y_i = \hat{f}(x)\}_{i=1}^N$. В практических случаях оценка функции плотности распределения по обучающей выборке зачастую существенно отличается от истинного распределения $p(x)$ на X.
4. Класс допустимых параметризованных аппроксиматоров $y = f(\theta, x): \Theta \otimes X \to \mathbb{R}^m$, где $\theta \in \Theta \subset \mathbb{R}^k$ – набор параметров аппроксимирующей функции, и Θ – компакт. В случае нейронной сети набор θ – веса нейронной сети.

Задача состоит в построении аппроксиматора $f^*(x)$ неизвестной функции $\hat{f}(x)$ на основании обучающей выборки S. Иными словами, $f^*(x)$ должна минимизировать интегральную ошибку (риск) аппроксиматора $f(\theta, x)$

$$loss = \int_X \|f(\theta, x) - \hat{f}(x)\| p(x) dx. \quad (1)$$

То есть $f^*(x) = f(\theta^*, x)$ для

$$\theta^* = \underset{\theta \in \Theta}{\operatorname{argmin}} \int_X \|f(\theta, x) - \hat{f}(x)\| p(x) dx. \quad (2)$$

При заданной конечной выборке $\{x_i\}_{i=1}^N$ назовем величину

$$loss_N = \frac{1}{N} \sum_{i=1}^N \|f(\theta, x) - y_i\| = \frac{1}{N} \sum_{i=1}^N \|f(\theta, x) - \hat{f}(x_i)\| \quad (3)$$

дискретной ошибкой аппроксиматора (эмпирическим риском) $f(\theta, x)$ на выборке $\{x_i\}_{i=1}^N$.

Из работы [12] известно, что

$$loss = \lim_{N \to \infty} loss_N \quad (4)$$

при условии, что $\{x_i\}_{i=1}^N$ – случайная последовательность с плотностью распределения $p(x)$ на X. Доказательству этого посвящены классические труды В.Н. Вапника и А.Я. Червоненкиса [12, 13].

Метод решения

На практике на основании того, что эмпирический риск сходится к риску при неограниченном росте размера выборки, и из-за того, что реальное распределение *P(x)* неизвестно, вместо общего риска используют среднеквадратичную ошибку (эмпирический риск). Зачастую количество доступных реальных сэмплов $\{x_i, y_i\}_{i=1}^N$ на много порядков меньше, чем доступное по вычислительным возможностям количество параметров. Так, характерным размером SOTA моделей распознавания образов являются модели

с числом параметров $10^9 \dots 10^{12}$. А характерные размеры размеченных датасетов реальных изображений общего вида составляют около $10^6 \dots 10^7$ изображений.

В практических кейсах на специфических доменах количество реальных изображений зачастую исчисляется тысячами или даже сотнями тысяч изображений. Например, в процессе оптимизации работы горно-обогатительного комбината (ГОКа) приходится на практике работать с малыми датасетами. В технологическом цикле ГОКа штатной аварийной ситуацией является наличие так называемого негабарита (фрагментов породы, линейные размеры которых больше размеров входного бункера дробильной машины) на конвейерной ленте. Необнаружение негабарита на конвейере приводит к забутовке дробилки и остановке всего технологического комплекса на многие часы, в течение которых проводятся работы по разрушению негабарита, тогда как обнаружение негабарита на конвейере задерживает работы на считанные минуты. Аналогичные проблемы возникают в задачах анализа степени автономности в робототехнических системах [14].

В этой ситуации стандартное решение задачи поиска оптимальной функции $f(\theta^*, x)$ путем оптимизации ошибки на выборке

$$\theta^* = \underset{\theta \in \Theta}{argmin} \frac{1}{N} \sum_{i=1}^{N} \| y_i - f(\theta, x_i) \| \tag{5}$$

ведет к переобучению.

Во введенных терминах переобучение означает, что алгоритм поиска оптимального θ в процессе своей работы достигает всё меньших значений эмпирического риска (уменьшается дискретная ошибка $loss_N$ аппроксиматора $f(\theta, x)$) с одновременным возрастанием полной ошибки аппроксиматора $loss$.

Оценим, насколько близки погрешность по выборке $\{x_i\}_{i=1}^N$ (эмпирический риск)

$$\frac{1}{N} \sum_{i=1}^{N} \| f(\theta, x_i) - y_i \| \tag{6}$$

и общий риск

$$\int_X \left\| f(\theta, x) - \hat{f}(x) \right\| p(x) dx. \tag{7}$$

Иначе говоря, оценим разницу

$$\left| \frac{1}{N} \sum_{i=1}^{N} \| f(\theta, x_i) - y_i \| - \int_X \left\| f(\theta, x) - \hat{f}(x) \right\| p(x) dx \right|. \tag{8}$$

Для этой цели рассмотрим стандартные $V_i = V(x_i)$ – области Вороного [15] для точек x_i в X:

$$V_i = \left\{ x \in X : \| x - x_i \| \leq \| x - x_j \| \ \ \forall \, j \neq i \right\}. \tag{9}$$

Технически области Вороного V_i и V_j с разными i и j могут иметь непустые пересечения. Мы будем пренебрегать этим на том основании, что интегралы рассматриваемых функций по пересечениям V_i и V_j всегда будут равны нулю. Поясним на примере функции

$$\int\limits_{V_i \cap V_j} \left\| f(\theta, x) - \hat{f}(x) \right\| p(x) dx.$$

В подынтегральном выражении $p(x)$ непрерывна на X, функции $f(\theta, x)$ и $\hat{f}(x)$ липшицевы, а значит, непрерывны на X. Операции взятия разности, нормы и произведения функций порождают непрерывные функции. Значит, всё подынтегральное выражение $\left\| f(\theta, x) - \hat{f}(x) \right\| p(x)$ – непрерывная функция. По определению ячеек Вороного пересечение V_i и V_j при $\forall\, j \neq i$ или пустое множество или, если V_i и V_j – «соседние» ячейки, $V_i \cap V_j$ представляет собой компактный фрагмент гиперплоскости (серединного перпендикуляра между x_i и x_j) с размерностью не больше $n-1$ в $X \subset \mathbb{R}^n$. Отсюда следует, что мера $V_i \cap V_j$ в $\mathbb{R}^n$ равна нулю, и интеграл непрерывной функции по $V_i \cap V_j$ тоже равен нулю.

Положим

$$m_i = m(V_i) = \int\limits_{V_i} p(x) dx. \tag{10}$$

Перепишем дискретную ошибку (эмпирический риск) – формулу (6) – в виде

$$\frac{1}{N} \sum_{i=1}^{N} \| f(\theta, x_i) - y_i \| = \frac{1}{N} \sum_{i=1}^{N} \| f(\theta, x_i) - \hat{f}(x_i) \|. \tag{11}$$

Подставим (11) в (8) и разобьем интеграл по X на сумму интегралов по областям Вороного V_i. Получаем

$$\left| \frac{1}{N} \sum_{i=1}^{N} \| f(\theta, x_i) - y_i \| - \int\limits_{X} \| f(\theta, x) - \hat{f}(x) \| p(x) dx \right| = \tag{12}$$

$$= \left| \frac{1}{N} \sum_{i=1}^{N} \| f(\theta, x_i) - \hat{f}(x_i) \| - \int\limits_{X} \| f(\theta, x) - \hat{f}(x) \| p(x) dx \right| = \tag{13}$$

$$= \left| \frac{1}{N} \sum_{i=1}^{N} \| f(\theta, x_i) - \hat{f}(x_i) \| - \sum_{i=1}^{N} \int\limits_{V_i} \| f(\theta, x) - \hat{f}(x) \| p(x) dx \right|. \tag{14}$$

Для дальнейшей оценки воспользуемся тем, что $\| f(\theta, x_i) - \hat{f}(x_i) \|$ не зависит от x, а

$$m_i = m(V_i) = \int\limits_{V_i} p(x) dx$$

по определению m_i. Отсюда следует, что

$$\sum_{i=1}^{N} \int_{V_i} \| f(\theta, x_i) - \hat{f}(x_i) \| p(x) dx = \sum_{i=1}^{N} m_i \| f(\theta, x_i) - \hat{f}(x_i) \|. \quad (15)$$

Добавим правую и левую часть этого выражения в (14) и сгруппируем слагаемые:

$$\left| \frac{1}{N} \sum_{i=1}^{N} \| f(\theta, x_i) - \hat{f}(x_i) \| - \sum_{i=1}^{N} \int_{V_i} \| f(\theta, x) - \hat{f}(x) \| p(x) dx \right| \leq$$

$$\leq \left| \frac{1}{N} \sum_{i=1}^{N} \| f(\theta, x_i) - \hat{f}(x_i) \| - \sum_{i=1}^{N} m_i \| f(\theta, x_i) - \hat{f}(x_i) \| \right| +$$

$$+ \left| \sum_{i=1}^{N} \int_{V_i} \| f(\theta, x_i) - \hat{f}(x_i) \| p(x) dx - \sum_{i=1}^{N} \int_{V_i} \| f(\theta, x) - \hat{f}(x) \| p(x) dx \right| \quad (15)$$

Оценим два последних слагаемых по отдельности.

Первое слагаемое в правой части неравенства (16) – это ошибка, обусловленная тем, что точки x_i недостаточно адекватно разбивают X на области Вороного и поэтому коэффициент $1/N$ не точно аппроксимирует меру соответствующей ячейки Вороного:

$$\left| \frac{1}{N} \sum_{i=1}^{N} \| f(\theta, x_i) - \hat{f}(x_i) \| - \sum_{i=1}^{N} m_i \| f(\theta, x_i) - \hat{f}(x_i) \| \right|. \quad (16)$$

Назовем ее ошибкой репрезентации.

Напомним, что

$$\sum_{i=1}^{N} m_i \| f(\theta, x_i) - \hat{f}(x_i) \| = \sum_{i=1}^{N} \int_{V_i} \| f(\theta, x_i) - \hat{f}(x_i) \| p(x) dx = \quad (17)$$

$$= \int_{X} \| f(\theta, x_i) - \hat{f}(x_i) \| p(x) dx, \quad (18)$$

потому что подынтегральное выражение не зависит от x, а m_i интеграл $p(x)$ по области Вороного.

Оценим этот вклад так:

$$\left| \frac{1}{N} \sum_{i=1}^{N} \| f(\theta, x_i) - \hat{f}(x_i) \| - \sum_{i=1}^{N} m_i \| f(\theta, x_i) - \hat{f}(x_i) \| \right| = \quad (19)$$

$$= \left| \sum_{i=1}^{N} \left(\frac{1}{N} - m_i \right) \| f(\theta, x_i) - \hat{f}(x_i) \| \right| \leq \quad (20)$$

$$\leq \sum_{i=1}^{N} \left|\frac{1}{N} - m_i\right| \|f(\theta, x_i) - \hat{f}(x_i)\|. \quad (21)$$

Дальнейшие оценки этого вклада могут опираться на то, насколько удачно выборка x_i представляет распределение $p(x)$.

Проиллюстрируем этот вклад на простейшем примере. Рассмотрим частный случай $n = 1$, $X = [0;1]$, а плотность распределения $p(x)$ непрерывна на X. Обозначим

$$a = \min_{x \in X} p(x),$$
$$b = \max_{x \in X} p(x).$$

Из условия $\int_X p(x)dx = 1$ и непрерывности $p(x)$ на X следует, что либо $a = b = 1$, либо $a < 1 < b$.

Напомним, что в исходной постановке задачи нам ничего неизвестно о распределении плотности $p(x)$. Требование непрерывности $p(x)$ всё ещё почти ничего не говорит о плотности распределения $p(x)$ на X. В отсутствие информации о распределении $p(x)$ рассмотрим такой конечный набор $\{x_i\}_1^N$, который будет наиболее репрезентативным одновременно для всех возможных распределений $p(x)$. Такой набор $\{x_i\}_1^N$ должен равномерно разбивать X на области Вороного. В нашем одномерном случае $n = 1$, $X = [0;1]$ – это отрезки $V_i = [i/N; (i+1)/N]$ для $i \in \{0,1,\dots,N-1\}$. Тем самым такой набор x_i воплощает наше незнание истинной функции плотности распределения $p(x)$.

Оценим сверху и снизу m_i. Подставляя в определение m_i оценки $0 < a \leq p(x) \leq b$, получаем:

$$\frac{a}{N} = \int_{V_i} a dx \leq m_i = m(V_i) = \int_{V_i} p(x)dx \leq \int_{V_i} b dx = \frac{b}{N}. \quad (22)$$

Учитывая, что $a \leq 1$ и $1 \leq b$ и оценку $a/N \leq m_i \leq b/N$, получаем:

$$\left|\frac{1}{N} - m_i\right| \leq \max\left\{\frac{1-a}{N}, \frac{b-1}{N}\right\} < \frac{b}{N}. \quad (23)$$

А значит, вклад этой компоненты в общую ошибку можно оценить через известный нам эмпирический риск так:

$$\sum_{i=1}^{N} \left|\frac{1}{N} - m_i\right| \|f(\theta, x_i) - \hat{f}(x_i)\| \leq b \sum_{i=1}^{N} \frac{1}{N} \|f(\theta, x_i) - \hat{f}(x_i)\| = b \cdot loss_N. \quad (24)$$

Этот пример иллюстрирует, как именно несоответствие фактической выборки x_i истинному (неизвестному) распределению $p(x)$ влияет на достижимую точность в оценке общего риска.

Второе слагаемое в правой части неравенства (16) – это ошибка, обусловленная тем, что в соответствующей области Вороного в формуле эмпирического риска используется $f(\theta, x_i)$ при фиксированном x_i, тогда как в формуле общего риска в подынтегральном выражении используется $f(\theta, x)$.

$$\left|\sum_{i=1}^{N}\int_{V_i}\|f(\theta,x_i)-\hat{f}(x_i)\|p(x)dx-\sum_{i=1}^{N}\int_{V_i}\|f(\theta,x)-\hat{f}(x)\|p(x)dx\right|. \quad (25)$$

Назовем ее ошибкой уклонения.

Рассмотрим случай, когда функции $f(\theta,x)$ и $\hat{f}(x)$ липшицевы, т.е.

$$\|f(\theta,x)-f(\theta,x')\|\le A|x-x'| \quad (26)$$

$$\|\hat{f}(x)-\hat{f}(x')\|\le B|x-x'| \quad (27)$$

$$\forall x,x'\in X,\theta\in\Theta \quad (28)$$

для некоторых $A > 0$ и $B > 0$. Тогда функция $\|f(\theta,x)-\hat{f}(x)\|$ тоже липшицева по x с константой $C = A + B$. Воспользуемся этим:

$$\left|\frac{1}{N}\sum_{i=1}^{N}\|f(\theta,x_i)-y_i\|-\int_{X}\|f(\theta,x)-\hat{f}(x)\|p(x)dx\right|= \quad (29)$$

$$\le\left|\sum_{i=1}^{N}\int_{V_i}\|f(\theta,x_i)-\hat{f}(x_i)\|-\|f(\theta,x)-\hat{f}(x)\|p(x)dx\right|\le \quad (30)$$

$$\le\sum_{i=1}^{N}\int_{V_i}\left|\|f(\theta,x_i)-\hat{f}(x_i)\|-\|f(\theta,x)-\hat{f}(x)\|\right|p(x)dx\le \quad (31)$$

$$\le\sum_{i=1}^{N}\int_{V_i}(A+B)\max_{x\in V_i}|x-x_i|p(x)dx= \quad (32)$$

$$=C\sum_{i=1}^{N}\max_{x\in V_i}|x-x_i|\int_{V_i}p(x)dx= \quad (33)$$

$$=C\sum_{i=1}^{N}m_i\max_{x\in V_i}|x-x_i|\le \quad (34)$$

$$\le C\max_{1\le i\le N}\max_{x\in V_i}|x-x_i|\sum_{i=1}^{N}m_i= \quad (35)$$

$$=C\max_{1\le i\le N}\max_{x\in V_i}|x-x_i|. \quad (36)$$

Последняя оценка фактически определяется диаметром максимальной области Вороного для выборки $\{x_i\}_1^N$. Сравнивая влияние двух факторов (19) и (36), можно сделать ряд выводов:

1. Ошибка репрезентации фактически определяется тем, насколько выборка x_i точно соответствует истинному распределению $p(x)$. В простейших случаях решающее значение имеет оценку сверху уклонения плотности распределения $p(x)$ на X от равномерного. Она может быть оценена через $loss_N$.

2. Ошибка уклонения фактически определяется константами Липшица функций $f(\theta, x)$ и $\hat{f}(x)$ и размерами областей Вороного V_i. Общепринято допущение, что x_i при стремлении N к бесконечности реализует распределение $p(x)$. В этом случае для почти всех выборок x_i решающий вклад в ошибку уклонения будет от ячеек Вороного в тех областях X, где $p(x)$ минимально, а значит, x_i в последовательности встречаются редко, и ячейки Вороного вокруг таких x_i имеют максимальный диаметр. С практической точки зрения это приводит к контринтуитивной рекомендации создавать выборки, уклоняющиеся в своем распределении от репрезентативных по $p(x)$.

На основе данных выводов рекомендуется создание обучающего набора данных в следующей последовательности:

1. На первоначальном, разведочном этапе нужно собрать набор $\{x_i\}_1^N$, который естественным образом репрезентирует распределение $p(x)$.
2. Так как истинное распределение $p(x)$ не известно, то следует построить приближенное распределение $p^*(x)$ на основании выборки $\{x_i\}_1^N$.
3. На базе $p^*(x)$ и равномерного на X распределения $p^0(x)$, используя параметр регуляризации α, построим распределение $p^{new}(x) = \alpha p^*(x) + (1-\alpha)\, p^0(x)$.
4. В соответствии с плотностью распределения $p^{new}(x)$ следует сгенерировать новые (синтетические или естественные) сэмплы x_i.

Выбор параметра регуляризации α должен опираться на оценки липшицевых констант. Чем они больше, тем более равномерно распределенной следует создавать обучающую выборку.

Заключение

В работе исследованы теоретические основания эффективности добавления в обучающую выборку дополнительных или синтетических данных.

В оценке общего риска выделено два типа ошибок: ошибка репрезентации и ошибка уклонения. Ошибка репрезентации обусловлена тем, что точки x_i недостаточно адекватно разбивают X на ячейки Вороного, и поэтому коэффициент усреднения $1/N$ неточно аппроксимирует меру соответствующей ячейки Вороного. Ошибка уклонения, обусловленная тем, что в соответствующей области Вороного в формуле эмпирического риска используется $f(\theta, x_i)$ при фиксированном x_i, тогда как в формуле общего риска в подынтегральном выражении используется $f(\theta, x)$.

Сформулированы практические рекомендации по созданию синтетических выборок, уклоняющихся в своем распределении от репрезентативных по $p(x)$, с

более частыми точками x_i в тех областях, где $p(x)$ имеет относительно малые значения для уменьшения размера соответствующих ячеек Вороного и уменьшения вклада ошибки уклонения в общий риск.

Библиографический список

1. Zhang H. et al. mixup: Beyond empirical risk minimization // arXiv preprint arXiv:1710.09412. 2017. https://doi.org/10.48550/arXiv.1710.09412.

2. Antoniou A., Storkey A., Edwards H. Data augmentation generative adversarial networks // arXiv preprint arXiv:1711.04340. 2017. https://doi.org/10.48550/arXiv.1711.04340.

3. Wu O., Li M. Revisiting the Effective Number Theory for Imbalanced Learning // IEEE Transactions on Knowledge & Data Engineering, vol. 36, no. 08, pp. 4192-4206, 2024. https://doi.org/10.1109/TKDE.2024.3367949.

4. Maharana K., Mondal S., Nemade B. A review: Data pre-processing and data augmentation techniques // Global Transitions Proceedings. — 2022. — Vol. 3, no. 1. — Pp. 91-99.

5. Yue Y., Li Y., Yi K., Wu Z. Synthetic Data Approach for Classification and Regression // 2018 IEEE 29th International Conference on Application-specific Systems, Architectures and Processors (ASAP). 2018. https://doi.org/10.1109/ASAP.2018.8445094.

6. El Emam K. et al. Utility metrics for evaluating synthetic health data generation methods: validation study // JMIR Medical Informatics. 2022. Vol. 10, no. 4. e35734. https://doi.org/10.2196/35734.

7. Borisov V., Leemann T., Seßler K., Haug J., Pawelczyk M., Kasneci G. Deep Neural Networks and Tabular Data: A Survey // IEEE Transactions on Neural Networks and Learning Systems. 2024. Vol. 35, no. 6. P. 7499-7519. https://doi.org/10.2196/3573410.1109/TNNLS.2022.3229161.

8. Douzas G., Lechleitner M., Bacao F. Improving the quality of predictive models in small data GSDOT: A new algorithm for generating synthetic data // PLoS ONE. 2022. Vol. 17, no. 4. e0265626. https://doi.org/10.1371/journal.pone.0265626.

9. Chatterjee S., Byun Y.-C. A Synthetic Data Generation Technique for Enhancement of Prediction Accuracy of Electric Vehicles Demand // Sensors. — 2023. — Vol. 23, no. 2. https://doi.org/10.3390/s23020594.

10. Iantovics L.B., Enăchescu C. Method for Data Quality Assessment of Synthetic Industrial Data // Sensors. 2022. Vol. 22, no. 4. 1608. https://doi.org/10.3390/s22041608.

11. Van Breugel B., Qian Z., van der Schaar M. Synthetic Data, Real Errors: How (Not) to Publish and Use Synthetic Data // Proceedings of the 40th International Conference on Machine Learning. 2023. Vol. 202.

12. Вапник В.Н., Червоненкис А.Я. О методе упорядоченной минимизации риска. I // Автоматика и телемеханика. 1974. № 8 С. 21-30.

13. Вапник В.Н., Червоненкис А.Я. О методе упорядоченной минимизации риска. II // Автоматика и телемеханика. 1974. № 9. С. 29-40.

14. Sokolov S., Sudakov, V. Multicriteria Analysis of the Robotic Systems Autonomy Using Fuzzy Calculations. In Proceedings of the 16th International Conference on Agents and Artificial Intelligence (ICAART 2024) — Volume 3, pp. 916-920. https://doi.org/10.5220/0012418200003636.

15. De Berg M., Cheong O., van Kreveld M., Overmars M. Computational Geometry: Algorithms and Applications. Springer Berlin, Heidelberg, 2008. 388 p. https://doi.org/10.1007/978-3-540-77974-2.

Оглавление