УДК 004.021 URL: https://ptsj.bmstu.ru/catalog/icec/inf_tech/1059.html

МЕТОД РАСШИРЕНИЯ ВЫБОРКИ ДЛЯ ОБУЧЕНИЯ МОДЕЛИ КЛАССИФИКАЦИИ НА ОСНОВЕ НАЛОЖЕНИЯ СЛУЧАЙНОГО ШУМА С УЧЕТОМ ЗНАЧЕНИЙ ЦЕЛЕВОГО ПРИЗНАКА

Н.П. Артюхин artyukhinnp@student.bmstu.ru
SPIN-код: 1465-5325

МГТУ им. Н.Э. Баумана, Москва, Российская Федерация

Исследована проблема недостаточного количества данных в выборке для обучения модели классификации и применения различных методов ее решения. Проанализирована предметная область данной проблемы и существующие методы увеличения размера обучающей выборки для модели классификации на основе двух подходов: добавление реальных данных и генерация искусственных данных. Сформулированы критерии сравнения данных методов. Разработан новый алгоритм увеличения размера обучающей выборки, которая состоит из структурированных данных, представленных в виде таблицы, на основе наложения случайного шума на числовые признаки и замены значений категориальных признаков наиболее часто встречающимися с учетом значения целевой переменной каждой записи исходной выборки. Исследовано влияние алгоритма увеличения размера выборки на качество модели классификации. Для этого проведено сравнение результатов обучения модели на исходной неувеличенной выборке, а также после применения каждого из рассмотренных методов расширения выборки: добавления реальных данных, добавления случайно сгенерированных данных, добавления перемешанных исходных данных, разработанного метода преобразования исходной выборки. Для оценки качества обученных моделей классификации использован коэффициент Джини. Показано, что в результате применения разработанного алгоритма к исходной обучающей выборке точность прогнозов модели классификации улучшилась и он превосходит аналогичные методы добавления синтетических данных.

Ключевые слова: машинное обучение, модель классификации, расширение обучающей выборки, преобразование данных, случайная генерация данных, перемешивание данных, структурированные данные, таблицы, коэффициент Джини

Введение. Для достижения высокой точности и надежности модели необходимо большое количество качественных данных. Их недостаток может привести к переобучению, когда модель довольно точно предсказывает классы объектов на обучающем наборе, но плохо обобщает их на новых данных [1].

Недостаточное количество данных может привести к неэффективному использованию вычислительных ресурсов и времени на обучение моделей, что также является важным аспектом в современных условиях. В некоторых областях, таких как медицинская или финансовая, сбор информации может

 ISSN 2541-8009

быть затруднен из-за этических или юридических ограничений [2]. По этой причине обостряется проблема нехватки данных для составления обучающих выборок.

В данной работе рассмотрены подходы к решению проблемы недостаточного количества данных в обучающих выборках, а именно алгоритмы генерации новых и преобразования исходных структурированных (табличных) данных для расширения выборок, использующихся для обучения моделей классификации.

Целью работы является создание метода увеличения размера обучающей выборки для модели классификации.

Методы увеличения выборки. Увеличение объема выборки имеет решающее значение для многих приложений искусственного интеллекта, поскольку точность увеличивается с увеличением объема обучающих данных [3]. Однако существует проблема, что у большинства компаний недостаточно данных для обучения своих моделей искусственного интеллекта. Именно здесь возникает необходимость расширения данных для их аналитики, прогнозов и рекомендаций, которые ранее были недоступны из-за недостаточного количества информации. Кроме того, использование небольшого объема данных может увеличить риск переобучения, в то время как наличие большего количества точек (примеров в выборке) позволяет его снизить [4]. Также необходимо учитывать, что данные для расширения выборки должны иметь тот же формат, что и исходные. При генерации новых данных или использовании нового источника информации необходимо обеспечить согласованность форматирования данных.

Существует два типа методов расширения выборки: увеличение выборки с помощью реальных данных и искусственное (синтетическое) увеличение выборки [4].

Реальное увеличение выборки. Реальное увеличение объема выборки осуществляется за счет объединения данных из других источников [5] или добавления новых атрибутов (признаков) в исходную выборку.

Достоинство данного метода — увеличение точности и надежности модели машинного обучения.

Недостатки данного метода:

– не всегда существует возможность применения данного метода (данных может быть мало, если область исследования новая);

– может потребоваться большое количество финансовых и временных ресурсов для получения реальных данных;

– данные из внешних источников могут отличаться по своему распределению от исходных данных.

Искусственное увеличение выборки. Существует два основных метода увеличения искусственного увеличения выборки: генерация случайных данных и преобразование существующих данных.

Генерация случайных данных. Это самый простой, но неэффективный метод повышения точности, поскольку при генерации искажается реальная зависимость целевой переменной (результата предсказания модели) от независимых переменных (признаков). Таким образом, вместо повышения точности предсказаний модели может произойти ее снижение.

Преобразование существующих данных. Основная идея метода заключается в том, чтобы создать новые, которые сохраняют основные характеристики исходных, но имеют некоторые отличия.

Для увеличения объема выборки изображений применяют геометрические преобразования (повороты изображений на заданный угол, отображение их по горизонтали или вертикали для изменения ориентации, удаление фрагментов изображения для фокусировки на определенных элементах или имитации более близкого обзора, перемещение изображений в различных направлениях) либо изменения цветовых свойств исходных изображений (изменение яркости изображения для имитации различных условий освещения, изменение контрастности, помогающее моделям распознавать объекты при различных уровнях четкости, изменение интенсивности цвета), зашумление изображений (гауссовы помехи, случайное добавление черных или белых пикселей) [6, 7].

Для увеличения объема выборки звуковых записей применяют наложение шума на исходные записи, изменяют частоту или амплитуду сигнала [8].

Для увеличения объема структурированных (табличных) данных иногда используют случайное наложение шума на числовые признаки или случайное перемешивание данных в столбцах, но это, вероятно, приведет к искажению реальных зависимостей между независимыми признаками и целевой переменной и, следовательно, к снижению качества прогнозов обученной модели. Для увеличения размера обучающей выборки можно также использовать алгоритм балансировки классов SMOTE (Synthetic Minority Over-Sampling Technique — метод избыточной выборки синтетического меньшинства), поскольку он создает новые синтетические данные, которые помогают модели лучше понять и обобщить характеристики минорных классов.

Алгоритм SMOTE состоит из трех этапов: 1) выбор образца из минорного класса в исходной выборке; 2) нахождение k ближайших к нему соседей (количество соседей k обычно выбирают экспериментально, слишком ма-

 ISSN 2541-8009

ленькое значение k может привести к недостаточному разнообразию, а слишком большое — к потере специфики малочисленного класса, обычно принимают $k = 5$) для создания данных, похожих на уже существующие, но при этом добавляющих разнообразия; 3) выбор одного из соседей случайным образом и интерполяция между ним и выбранным образцом минорного класса из исходной выборки для создания нового образца данного класса [9]. Интерполяция между образцами позволяет создавать новые данные, которые сохраняют общие характеристики малочисленного класса, но при этом добавляют немного разнообразия [10].

Других конкретных подходов к увеличению объема структурированных (табличных) данных в открытых источниках найдено не было. В случае табличных данных можно применять преобразование существующих строк таблицы, чтобы получить новые. Чтобы не исказить реальную зависимость целевой переменной (результата предсказания модели) от независимых переменных (признаков), необходимо проводить статистический анализ исходных данных.

Далее будут использованы следующие обозначения методов увеличения выборки:

– РУВ — реальное увеличение выборки;

– ГСД — генерация случайных данных;

– ПСД — преобразование существующих данных.

Основными характеристиками, определяющими эффективность применения данных методов, являются количество финансовых и временных ресурсов, сложность реализации, возможность применения данного подхода в любой ситуации (универсальность) и влияние на искажение реальной зависимости целевой переменной от признаков. На основе этого были выделены следующие критерии для оценки качества описанных методов:

– тип данных, используемых в методе (реальные или искусственные);

– затраты временных ресурсов;

– затраты денежных ресурсов;

– сложность реализации;

– универсальность метода;

– искажение реальной зависимости.

Результаты сравнения методов увеличения выборки по сформулированным критериям приведены в таблице.

Таким образом, на основании выделенных критериев было принято решение при разработке собственного метода увеличения размера обучающей выборки основываться на подходе преобразования существующих данных.

Сравнение методов увеличения выборки

Критерий сравнения	РУВ	ГСД	ПСД
Тип данных	Реальные	Искусственные	Искусственные
Затраты временных ресурсов	Высокие	Низкие	Средние
Затраты денежных ресурсов	Высокие	Низкие	Низкие
Сложность реализации	Низкая	Низкая	Высокая
Универсальность метода	Низкая	Высокая	Высокая
Искажение реальной зависимости	Низкое	Высокое	Среднее

Разработанный алгоритм увеличения размера выборки. Для того чтобы при преобразовании существующих данных для генерации новых не было искажения реальной зависимости целевой переменной от значений признаков, используется статистический анализ. Новые строки генерируются на основе каждой строки исходной выборки, т. е. размер увеличенной выборки станет в 2 раза больше размера исходной за одну итерацию алгоритма увеличения выборки. При определении значения каждого столбца новой строки учитываются все значения соответствующего столбца исходной выборки.

У новой строки сохраняется значение целевой переменной, т. е. берется ее значение в текущей преобразуемой строке исходной выборки. Значения числовых столбцов новой строки (целые/вещественные) получаются путем либо увеличения, либо уменьшения на [0, 100) процентов от исходного значения преобразуемой строки. Принятие решения об увеличении или уменьшении значения зависит от количества строк, у которых значение в данном столбце больше/меньше и значение целевой переменной такое же. Если количество строк, у которых значение в данном столбце больше, чем в текущей преобразуемой строке, и значение целевой переменной такое же, больше, чем количество строк, у которых значение в данном столбце меньше, чем в текущей преобразуемой строке, и значение целевой переменной такое же, то принимается решение об увеличении значения в текущем столбце преобразуемой строки, иначе — об уменьшении. Значения категориальных переменных новой строки получаются путем замены исходного значения на самое часто встречающееся, но отличное от текущего значение данной переменной, которое соответствует такому же значению целевой переменной, как у преобразуемой строки. Даты и логические переменные (признаки) не изменяются, то есть берутся из текущей преобразуемой строки. Таким образом, увеличивает-

 ISSN 2541-8009

ся вероятность того, что новая строка (новый элемент выборки) не исказит реальную зависимость между целевой переменной и значениями признаков.

Схема алгоритма увеличения выборки представлена на рис. 1.

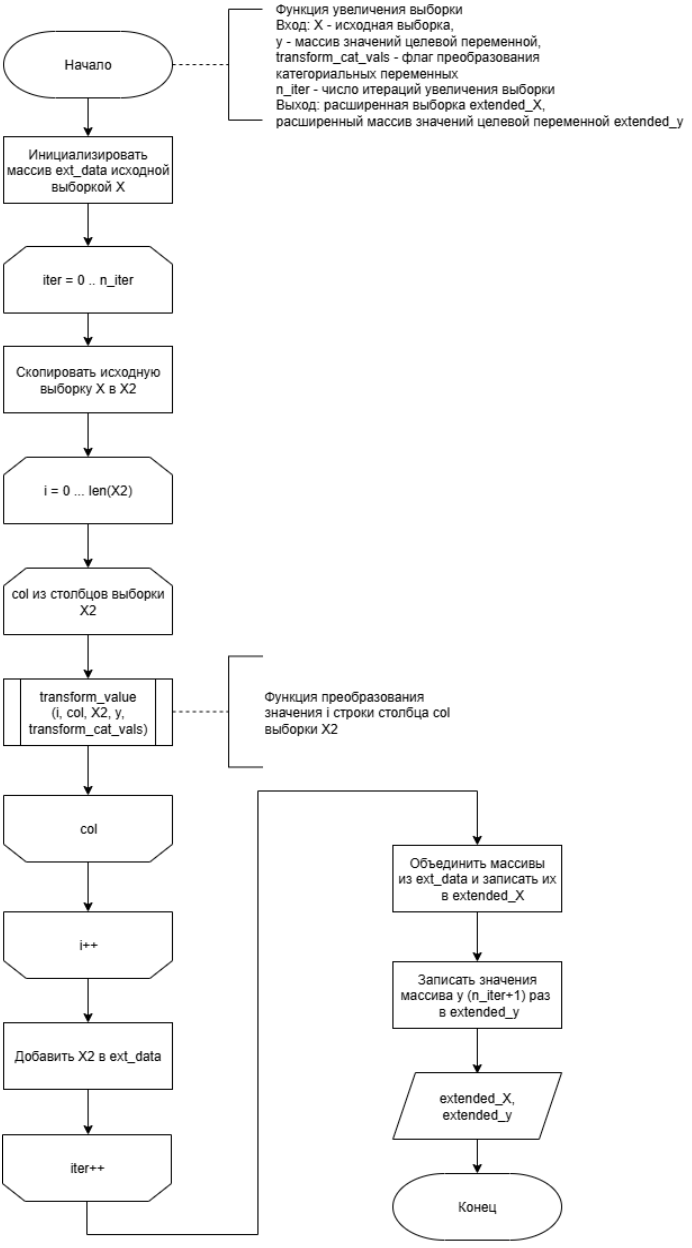

Рис. 1. Схема алгоритма увеличения выборки

Исследование влияния выбора алгоритма на качество прогнозирования модели классификации. Исходная обучающая выборка составлена для решения задачи разделения кредитных заявок на две группы: «0» — с низкой вероятностью дефолта (рекомендуются к одобрению), «1» — с высокой вероятностью дефолта (не рекомендуются к одобрению), имеет размер 100 строк (50 записей со значением целевой переменной, равной единице, и 50 записей со значением целевой переменной, равной нулю). Размер выборки варьируется от 100 до 500 записей (исходная выборка увеличивается в 2, 3, 4, 5 раз). Количество признаков в исходной выборке — 280. Среди них данные из кредитной заявки клиента (сумма кредита, срок кредита, дата заявки, город заявки, заявленный доход за месяц, заявленные средние расходы за месяц и т. п.) и данные из бюро кредитных историй (количество и суммы действующих, погашенных и просроченных потребительских кредитов, автокредитов, POS-кредитов, ипотек).

Увеличение размера выборки осуществляется четырьмя способами, а именно за счет добавления:

1) реальных данных;

2) искусственных (сгенерированных) данных, полученных с помощью разработанного алгоритма увеличения выборки, преобразующего исходные данные и учитывающего значения целевой переменной (target);

3) искусственных (сгенерированных) данных, полученных с помощью алгоритма, случайно преобразующего исходные данные и не учитывающего значения целевой переменной (target);

4) искусственных (сгенерированных) данных, полученных с помощью случайного перемешивания данных в столбцах исходной выборки.

Применение алгоритма SMOTE не рассматривалось, поскольку данный алгоритм может увеличить данные только за счет балансировки классов, в результате его применения количество данных в выборке может быть увеличено менее чем в 2 раза.

Фрагмент исходной выборки из 10 записей (со следующими признаками: bank_city — город кредитной заявки, pledge — признак наличия залога, req_amount — запрашиваемая сумма кредита, req_term – запрашиваемый срок погашения кредита, req_income – заявленный доход в месяц в рублях, req_outcome — заявленные средние расходы за месяц в рублях, business_owner_my — признак наличия своего бизнеса у заемщика, app_id — ID заявки, target — признак дефолта заемщика) и результат применения к нему разработанного алгоритма увеличения выборки (число записей было увеличено в 2 раза: с 10 до 20) представлены на рис. 2, 3. На рис. 3 строки с номерами 2–11 являются строками исходной выборки, к которой применялся разра-

 ISSN 2541-8009

ботанный алгоритм, а строки с номерами 12–21 были сгенерированы данным алгоритмом. На основе изменения строки 2 алгоритм получил строку 12 (у строки 2 признак дефолта равен нулю, т. е. дефолт не случился, поэтому у строки 12 алгоритм также проставил признак дефолта ноль; город заявки поменялся с Уфы на Ярославль, так как в исходной выборке есть заявка из Ярославля с признаком дефолта, равным нулю; признак наличия залога остался равным нулю, так как в исходной выборке больше кредитных заявок без дефолта с признаком залога равным 0; сумма кредита была увеличена с 240 до 470 тысяч рублей, так как в исходной выборке у большинства кредитных заявок без дефолта сумма кредита выше, чем в строке 2; срок кредита увеличен с 48 до 56, так как у большинства кредитных заявок без дефолта в исходной выборке срок кредита больше, чем в строке 2, и т. д. по каждому столбцу выборки), на основе строки 3 — строку 13, на 4 — 14 и т. д.

	bank_city	pledge	req_amount	req_term	req_income	req_outcome	business_owner_my	app_id	target
2	уфа	0	240000.0	48	60000.0	0.0	0	1266912	0
3	казань	0	300000.0	30	125000.0	45000.0	0	1267318	0
4	краснодар	1	2200000.0	60	0.0	168015.22	1	1267355	0
5	ярославль	0	200000.0	24	48000.0	5200.0	0	1268047	0
6	белгород	0	250000.0	60	95000.0	0.0	0	1268169	1
7	санкт-петербург	1	1500000.0	36	0.0	194454.21	1	1268315	0
8	ростов-на-дону	0	250000.0	60	0.0	0.0	0	1268441	0
9	иваново	0	200000.0	60	84500.0	27000.0	0	1268597	1
10	воронеж	0	250000.0	60	0.0	0.0	0	1268943	1
11	москва	0	750000.0	60	0.0	123590.01	1	1269072	1

Рис. 2. Фрагмент исходной выборки из 10 записей

	bank_city	pledge	req_amount	req_term	req_income	req_outcome	business_owner_my	app_id	target
2	уфа	0	240000.0	48	60000.0	0.0	0	1266912	0
3	казань	0	300000.0	30	125000.0	45000.0	0	1267318	0
4	краснодар	1	2200000.0	60	0.0	168015.22	1	1267355	0
5	ярославль	0	200000.0	24	48000.0	5200.0	0	1268047	0
6	белгород	0	250000.0	60	95000.0	0.0	0	1268169	1
7	санкт-петербург	1	1500000.0	36	0.0	194454.21	1	1268315	0
8	ростов-на-дону	0	250000.0	60	0.0	0.0	0	1268441	0
9	иваново	0	200000.0	60	84500.0	27000.0	0	1268597	1
10	воронеж	0	250000.0	60	0.0	0.0	0	1268943	1
11	москва	0	750000.0	60	0.0	123590.01	1	1269072	1
12	ярославль	0	469703.16114824126	56	50668.04922050029	0.9586334004094057	0	2332595	0
13	ярославль	0	405049.640054929	39	2853.168997337634	0.0	0	2539290	0
14	ярославль	1	385169.760050921	34	0.8091990493147914	57312.02839816053	1	2058570	0
15	ростов-на-дону	0	323889.37522019306	38	11937.395917376634	97.47642030089628	0	2520850	0
16	иваново	0	56541.35624424441	31	75870.75499410639	0.6889389323738385	0	1424058	1
17	ярославль	1	1363155.0935129162	55	0.8244611016881447	162155.2207923987	1	1972138	0
18	ярославль	0	290044.61608447036	45	0.8311657510551862	0.5822916789118604	0	2382648	0
19	воронеж	0	276856.97092234856	54	4002.400282169532	1659.6775467885964	0	2466970	1
20	иваново	0	389047.3579603224	11	0.9838362660876219	0.8124852838559621	0	1636891	1
21	иваново	0	665651.1555416388	17	0.7304054781801408	102781.653316567	1	1897046	1

Рис. 3. Результат применения разработанного алгоритма увеличения размера выборки

Оценка качества модели классификации кредитных заявок, разработанной для проведения данного исследования, по значению коэффициента Джини [11, 12] была выполнена на тестовой выборке размером 8000 записей. Поскольку при генерации искусственных данных используются случайные числа, при каждом запуске алгоритма искусственного увеличения выборки результат его работы будет разным и, следовательно, будет варьироваться точность прогнозов обученной модели. Поэтому при исследовании алгоритмы искусственного увеличения размера выборки запускались по 100 раз и учитывалось максимальное значение коэффициента Джини обученной на расширенной выборке модели. График зависимости значения коэффициента Джини от кратности увеличения исходной выборки с помощью рассмотренных алгоритмов приведен на рис. 4.

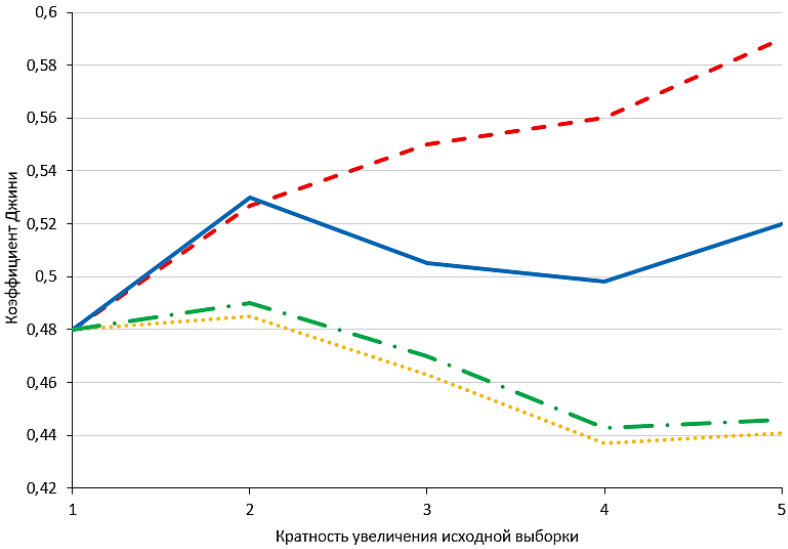

Рис. 4. График зависимости значения коэффициента Джини от кратности увеличения исходной выборки:
линия *красного* цвета (штрих) — увеличение за счет добавления реальных данных; линия *синего* цвета (сплошная) — искусственное увеличение (с учетом целевой переменной); линия *желтого* цвета (пунктир) — искусственное увеличение (без учета целевой переменной); линия *зеленого* цвета (штрих-пунктир) — искусственное увеличение перемешиванием

В результате проведенного исследования было получено, что увеличение размера выборки за счет добавления реальных данных представляет собой наиболее эффективный метод повышения качества модели прогнозирования, однако далеко не всегда есть такая возможность. Чем больше реальных дан-

 ISSN 2541-8009

ных добавлено в исходную выборку, тем более высоким становится значение коэффициента Джини, т. е. обученная модель дает более точные прогнозы.

Разработанный алгоритм искусственного увеличения размера выборки, который учитывает значения целевой переменной повышает качество модели прогнозирования на 10…20 % больше, чем другие алгоритмы искусственного увеличения выборки. Максимальный прирост коэффициента Джини за счет добавления искусственных данных достигается при увеличении объема исходной выборки в 2 раза, при дальнейшем увеличении объема возможно снижение качества модели из-за того, что доля реальных данных в выборке снижается и повышается риск искажения реальной зависимости целевой переменной от независимых переменных (признаков).

Заключение. В результате проделанной работы получены следующие результаты.

1. На основе анализа предметной области проблемы недостаточного количества данных были выделены следующие методы расширения обучающих выборок (табличных данных): добавление реальных данных из внешних источников, генерация новых случайных данных, преобразование данных исходной выборки за счет наложения случайного шума на числовые признаки, добавление перемешанных данных из исходной выборки, алгоритм балансировки классов SMOTE.

2. Сформулированы следующие критерии сравнения существующих подходов к увеличению размера обучающей выборки: затраты временных ресурсов, затраты финансовых ресурсов, сложность реализации, универсальность метода, влияние на искажение реальной зависимости целевой переменной от независимых признаков. По данным критериям выявлен наиболее оптимальный подход к расширению выборки — преобразование исходных данных.

3. Разработан новый алгоритм увеличения объема выборки (табличных данных) для модели классификации на основе наложения случайного шума на числовые признаки и замены значений категориальных признаков модой с учетом значений целевой переменной исходного набора данных.

4. В результате сравнения разработанного алгоритма расширения обучающей выборки с рассмотренными было установлено, что он на 10…20 % эффективнее повышает качество модели классификации, оцениваемое по значению коэффициента Джини, чем другие рассмотренные алгоритмы генерации искусственных данных. Наилучший результат разработанный алгоритм показывает при увеличении размера исходной выборки в 2 раза, далее может наблюдаться снижение качества обученной модели классификации из-за преобладания искусственных данных над реальными.

Литература

[1] Mumuni A., Mumuni F. Data augmentation: A comprehensive survey of modern approaches. *Array*, 2022, vol. 16 (6), art. no. 100258. https://doi.org/10.1016/j.array.2022.100258

[2] Мельникова М.Е. Порядок и условия обработки персональных данных. *National Science Journal*, 2022, № 1, с. 16–21.

[3] Fonseca J., Bacao F. *Research Trends and Applications of Data Augmentation Algorithms*. NOVA Information Management School, Universidade Nova de Lisboa, 2022. https://doi.org/10.48550/arXiv.2207.08817

[4] *Data Augmentation for Machine Learning*. URL: https://www.akkio.com/data-augmentation-for-machine-learning (accessed 19.10.2024).

[5] *Open Source Data Repositories for ML*. URL: https://www.restack.io/p/ci-cd-machine-learning-answer-open-source-data-repositories-cat-ai (accessed 22.10.2024).

[6] Shorten C., Khoshgoftaar T.M. A survey on Image Data Augmentation for Deep Learning. *Journal of Big Data*, 2019, vol. 6 (1). https://doi.org/10.1186/s40537-019-0197-0

[7] Alomar K., Aysel H.I. Data Augmentation in Classification and Segmentation: A Survey and New Strategies. *Journal of Imaging*, 2023, vol. 9 (2), art. no. 46. https://doi.org/10.3390/jimaging9020046

[8] Wei S., Zou S., Liao F. A Comparison on Data Augmentation Methods Based on Deep Learning for Audio Classification. *Journal of Physics Conference Series*, 2020, vol. 1453 (1), art. no. 012085. https://doi.org/10.1088/1742-6596/1453/1/012085

[9] Blagus R., Lusa L. SMOTE for High-Dimensional Class-Imbalanced Data. *BMC Bioinformatics*, 2013, vol. 106. https://doi.org/10.1186/1471-2105-14-106

[10] Nitesh V.C., Kevin W.B., Lawrence O.H. et al. SMOTE: synthetic minority oversampling technique. *Journal of Artificial Intelligence Research*, 2002, vol. 16 (1), pp. 321–357. https://doi.org/10.1613/jair.953

[11] *Gini Index: Decision Tree, Formula, Calculator, Gini Coefficient in Machine Learning*. URL: https://blog.quantinsti.com/gini-index/ (accessed 02.11.2024).

[12] Farris F.A. The Gini Index and Measures of Inequality. *The American Mathematical Monthly*, 2010, vol. 117 (10), pp. 851–864. https://doi.org/10.4169/000298910X523344

Поступила в редакцию 16.02.2025

 ISSN 2541-8009

Артюхин Николай Павлович — студент магистратуры кафедры «Программное обеспечение ЭВМ и информационные технологии», МГТУ им. Н.Э. Баумана, Москва, Российская Федерация.

Научный руководитель — Барышникова Марина Юрьевна, кандидат педагогических наук, доцент кафедры «Программное обеспечение ЭВМ и информационные технологии», МГТУ им. Н.Э. Баумана, Москва, Российская Федерация. E-mail: baryshnikovam@bmstu.ru, SPIN-код: 9019-8093.

Ссылку на эту статью просим оформлять следующим образом:
Артюхин Н.П. Метод расширения выборки для обучения модели классификации на основе наложения случайного шума с учетом значений целевого признака. *Политехнический молодежный журнал*, 2025, № 04 (99). URL: https://ptsj.bmstu.ru/catalog/icec/inf_tech/1059.html

DATA AUGMENTATION METHOD FOR TRAINING A CLASSIFICATION MODEL BASED ON THE IMPOSITION OF RANDOM NOISE, TAKING INTO ACCOUNT THE VALUES OF THE TARGET FEATURE

N.P. Artyukhin artyukhinnp@student.bmstu.ru
SPIN-code: 1465-5325

Bauman Moscow State Technical University, Moscow, Russian Federation

The paper investigates the problem of insufficient data in the sample to train the classification model and application of various methods to solve it. It analyzes the subject area devoted to this problem and the existing data augmentation methods for the classification model. It formulates the criteria for comparing these methods. The paper describes a new data augmentation algorithm based on the imposition of random noise on numerical features and replacing the values of categorical features with the most common ones, taking into account the value of the target feature of each record of the original dataset. It studies the impact of the developed data augmentation algorithm on the quality of the classification model. To do this, it compares the results of training the model on the initial dataset and after applying each of the considered methods of sample expansion: adding real data, adding randomly generated data, adding mixed initial data. It uses the Gini coefficient to assess the quality of the trained classification models. It shows that the result of applying the developed algorithm to the initial training dataset lead to improving the accuracy of the classification model predictions and developed algorithm surpasses similar methods of adding synthetic data.

Keywords: machine learning, classification model, training dataset augmentation, data transformation, random data generation, data mixing, structured data, tables, Gini index

Received 16.02.2025

Artyukhin N.P. — Magistracy Student of Department of Computer Software and Information Technology, Bauman Moscow State Technical University, Moscow, Russian Federation.

Scientific advisor — Baryshnikova M.Yu., Ph. D. (Ped.), Associate Professor of Department of Computer Software and Information Technology, Bauman Moscow State Technical University, Moscow, Russian Federation. E-mail: baryshnikovam@bmstu.ru, SPIN-code: 9019-8093.

Please cite this article in English as:
Artyukhin N.P. Data augmentation method for training a classification model based on the imposition of random noise, taking into account the values of the target feature. *Politekhnicheskiy molodezhnyy zhurnal*, 2025, no. 04 (99). (In Russ.). URL: https://ptsj.bmstu.ru/catalog/icec/inf_tech/1059.html

 ISSN 2541-8009