Оригинальная статья
https://doi.org/10.21683/1729-2646-2021-21-3-54-65

Интеллектуальные методы повышения точности прогнозирования редких опасных событий на железнодорожном транспорте

Ольга Б. Проневич[1*], Михаил В. Зайцев[1]
[1]АО «НИИАС», Москва, Российская Федерация
**obpronevich@gmail.com*


Ольга Б. Проневич


Михаил В. Зайцев

Резюме. Цель статьи – *рассмотреть подходы к методам повышения качества прогнозирования и классификации несбалансированных данных и выбрать методы, позволяющие повысить точность классификации редких событий. При прогнозировании появления редких событий методами машинного обучения ученые сталкиваются с проблемой несоответствия качества обученных моделей их реальной способности правильно спрогнозировать появление редкого события. Предмет исследования в статье – обучение моделей при исходных несбалансированных данных. Объект исследования – информация об инцидентах и опасных событиях на объектах железнодорожного электроснабжения. Проблема несбалансированных данных выражается заметной диспропорции между типами наблюдаемых событий – количествами представителей различных классов.* **Методы.** *При работе с несбалансированными данными, в зависимости от характера задачи, качества и объема исходных данных, применяют различные методы повышения качества моделей классификации и прогнозирования Data Science. Часть этих методов направлена на работу с признаками и параметрами моделей классификации. К ним относятся методы FAST, CFS, нечёткие классификаторы, GridSearchCV и другие. Другая группа методов ориентирована на формирование репрезентативных подмножеств из исходного массив данных – сэмплов. Методы сэмплинга данных позволяют исследовать влияние пропорции классов на качество машинного обучения. В частности, в рамках настоящей статьи подробно рассматривается метод NearMiss.* **Результаты.** *Проблема дисбаланса классов при анализе количества инцидентов на объектах железнодорожного транспорта существуют с 2015 года. Несмотря на снижение доли опасных событий на объектах железнодорожного электроснабжения в течении трех лет с 2018 года, не исключен рост количества таких событий. Статистика долей опасных событий на уровне месяца демонстрирует отсутствие тренда на снижение и наличие пиков. В таких условиях эффективным периодом наблюдений за количеством инцидентов и опасных событий является месяц. Визуализация соотношения классов показала отсутствие выраженной границы между представителями класса большинства (инцидентами) и класса меньшинства (опасные события). Исследовалось соотношение классов в двух и трех измерениях в натуральных величинах и с применением метода главных компонент. Такая «близость» классов является одной из причин ошибок прогноза. В рамках работы проведен анализ имеющегося исследовательского опыта повышения качества машинного обучения при работе с несбалансированными данными. Определены и уточнены используемые для описания степени дисбалансов классов термины. Изучены сильные и слабые стороны различных методов работы с такими данными, приведено описание сильных и слабых сторон 50 методов. Из методов работы с количеством представителей классов при решении задачи классификации (прогнозирования появления) редких опасных событий на железнодорожном транспорте выбран метод NearMiss. Указанный метод позволяет проводить эксперименты с пропорциями представителей классов и методами отбора представителей классов. По результатам серии экспериментов удалось добиться повышения точности классификации редких опасных событий от 0 до 70-90%.*

Ключевые слова: *машинное обучение, редкие события, дисбаланс классов, повышение точности прогнозирования, сэмплинг данных, балансировка классов.*

Для цитирования: *Проневич О.Б., Зайцев М.В. Интеллектуальные методы повышения точности прогнозирования редких опасных событий на железнодорожном транспорте // Надежность. 2021. №3. С. 54-64. https://doi.org/10.21683/1729-2646-2021-21-3-54-64*

Поступила *05.07.2021 г.* / **После доработки** *02.08.2021 г.* / **К печати** *17.09.2021 г.*

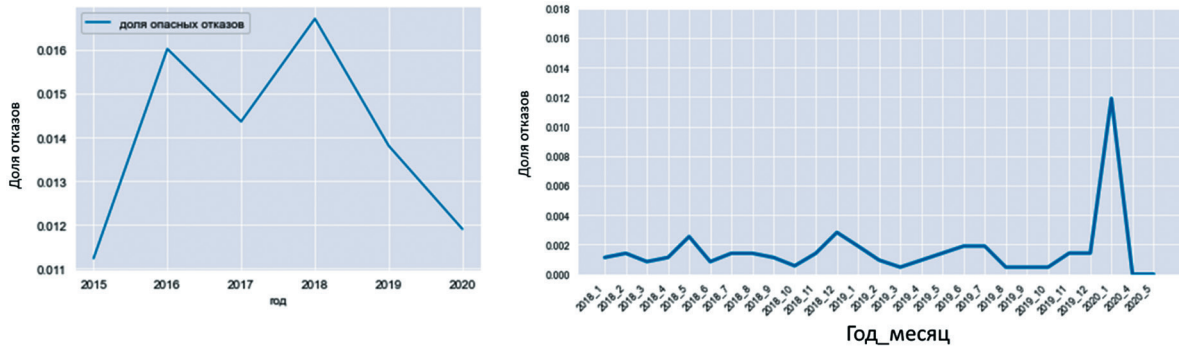
1. Актуальность проблемы повышения качества моделирования при прогнозировании редких событий на железнодорожном транспорте

При прогнозировании опасных событий исследователи сталкиваются с противоречивой проблемой: чем меньше событий, тем лучше для объекта наблюдения, и тем сложнее обучить модель необходимого качества. В настоящее время доля опасных событий из всех инцидентов с объектами железнодорожного электроснабжения (ЖЭс) не превышает 2% в год (рис. 1а).

Как видно из графика на рис. 1а, доля опасных событий с 2018 года падает, однако доступного периода наблюдения недостаточно для того, чтобы сделать выводы о стабильном снижении, а данные до 2018 года свидетельствуют о возможности роста количества отказов. Детализация статистики до уровня месяца (рис. 1б) показывает наличие пиков количества опасных событий. Отсутствие выраженной сезонности в данных не позволяет планировать периодические профилактические мероприятия для предотвращения появления опасных событий. Год, как период наблюдения, не является эффективным, т.к. состояние объектов железнодорожного транспорта за год эксплуатации сильно меняется и запланированные работы могут оказаться неактуальными. Поэтому особый интерес представляют модели, которые позволяют прогнозировать появление опасных событий по месяцам. На этом уровне детализации наблюдений вопрос «редкости» целевого фактора встает еще острее – в среднем на него приходится менее чем 0,4% случаев.

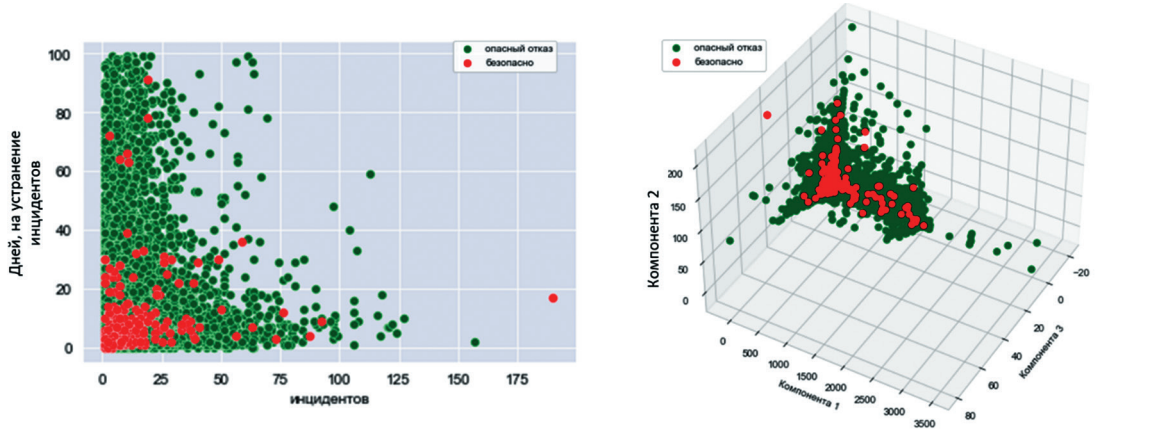
Несмотря на то, что количество опасных событий мало, количество не относящихся к ним инцидентов измеряется тысячами, что позволяет применять методы интеллектуального анализа больших данных. На пути построения высокоточных моделей существуют два основных препятствия: дисбаланс классов и качество данных. Когда говорят о качестве данных, обычно имеют в виду такие характеристики, как: неполнота, дублирование, несогласующиеся друг с другом сведения, ошибки при ручном вводе данных. Однако при работе с объектами железнодорожного транспорта авторы столкнулись с другой проблемой, а именно – слабыми отличиями между характеристиками объектов. Основная информация, характеризующая объекты ЖЭс на уровне месяца – данные об инцидентах. После подготовки данных каждый объект ЖЭс характеризуется более чем 150 признаками.

На рис. 2а показано соотношение классов на плоскости с двумя координатами: количество инцидентов и среднее количество дней на устранение инцидентов.



а) доля опасных событий по годам б) доля опасных событий по месяцам

Рис. 1. Доля опасных событий в инцидентах с объектами ЖЭс



а) Дисбаланс классов на плоскости б) дисбаланс классов в 3-х измерениях

Рис. 2. Дисбаланс классов на объектах со схожими характеристиками

Видно, что для этих двух признаков нет границы между классами. Применив метод главных компонент для уменьшения размерности, перейдем от сотни признаков к трем (рис. 2б). На графике лучше видна концентрация опасных событий, однако по-прежнему нет выраженной границы между классами.

Таким образом, путь повышения качества классификации инцидентов должен включать в себя работу с дисбалансом классов и отбором признаков. Без работы по указанным направлениям качество моделей классификации состояния объектов ЖЭс не будет удовлетворительным. (табл. 1).

Табл. 1. Качество моделей классификации до применения методов повышения точности

Модель	Точность прогноза опасного события		Точность прогноза безопасного состояния	
	Обучающая выборка	Тестовая выборка	Обучающая выборка	Тестовая выборка
GBC	0	0,416	1	1
Logi	0	0	1	1
KNN	0	0	1	1
DTC	0	0,043	0,994	1

2. Направления повышения точности предсказания появления редких событий

Методы повышения качества классификации можно разделить на две группы: работа с признаками и параметрами, а также работа с количеством представителей классов. Эти методы могут быть использованы как на этапе подготовки данных, так и на этапе обучения (рис. 3).

К группе методов работы с признаками относятся:

1) методы отбора значимых признаков;

2) выбор и настройка моделей.

Методы работы с количеством представителей классов:

1) уравнивание классов;

2) работа с дисбалансом классов;

3) моделирование на классе меньшинства;

4) подбор эффективной функции штрафов (работа с ошибкой классификации).

Несмотря на то, что методы в настоящее время широко известны, не существует единого алгоритма их комбинированного использования для повышения качества классификации. Специфика и объем данных, доступный при решении каждой конкретной задачи, вынуждает исследователей проводить эксперименты, в поисках оптимального сочетания различных методов. Так, например, в работе [1] предложена схема, состоящая из использования комбинации алгоритмов классификации и методов отбора признаков *RFE*, *Random Forest* и *Boruta*, с предварительным использованием балансирования классов методами случайного сэмплирования *SMOTE* и *ADASYN*. Эта работа показывает повышение точности классификации до 98% (с 93%). Однако исследование [2], посвященное классификации текстов, показывает, что эффективнее работать с изменением порога решений и весами ошибок разного рода, чем пытаться изменить распределение. При этом автор вводит разделение несбалансированных данных на умеренно несбалансированные данные (соотношение классов 7 к 1) и сильно несбалансированные данные (соотношение классов 14 к 1). В статье [2] продемонстрировано, что работа над пропорциями классов на сильно несбалансированных данных для некоторых моделей приводит к повышению точности классификации. При классификации обращений пользователей [3] определено, что применение методов балансирования классов может не только не дать результат, но и привести к снижению точности классификации. Вопреки исследованию [3], в статье [4] продемонстрирована эффективность применения методов сэмплинга данных.

Цель настоящей статьи – продемонстрировать на практике влияние методов работы с представителями классов на точность классификации опасных событий с объектами ЖЭс.

2.1. Обзор методов работы с дисбалансом классов.

Одним из основных методов работы с дисбалансом классов является метод *SMOTE*, алгоритм которого был разработан в начале 2002 г. [5]. В настоящее время су-

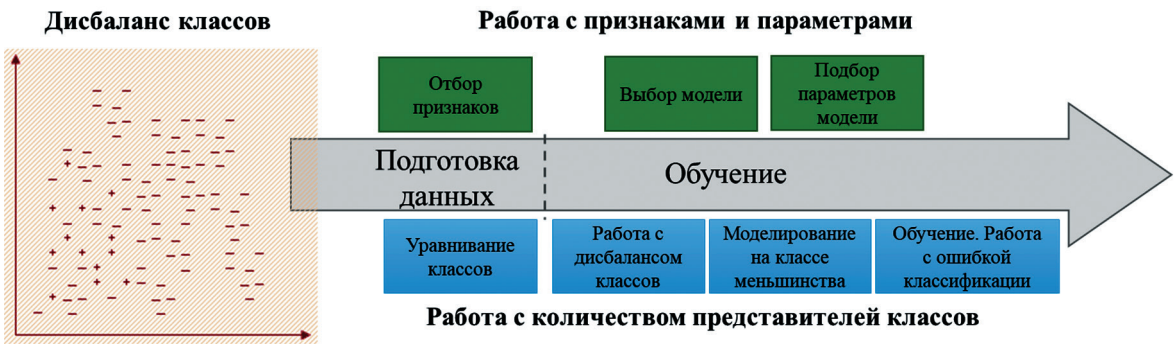


Рис. 3. Пути повышения качества классификации

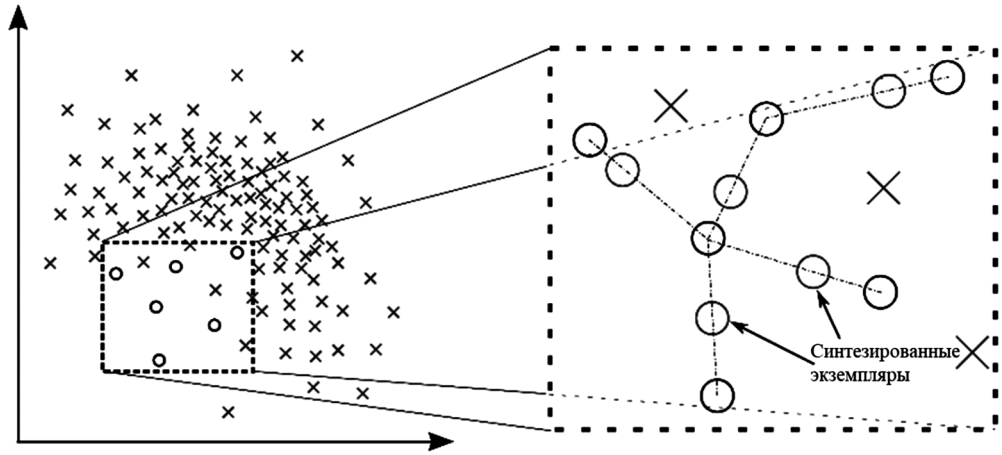


Рис. 4. Синтезирование наблюдения методом *SMOTE*

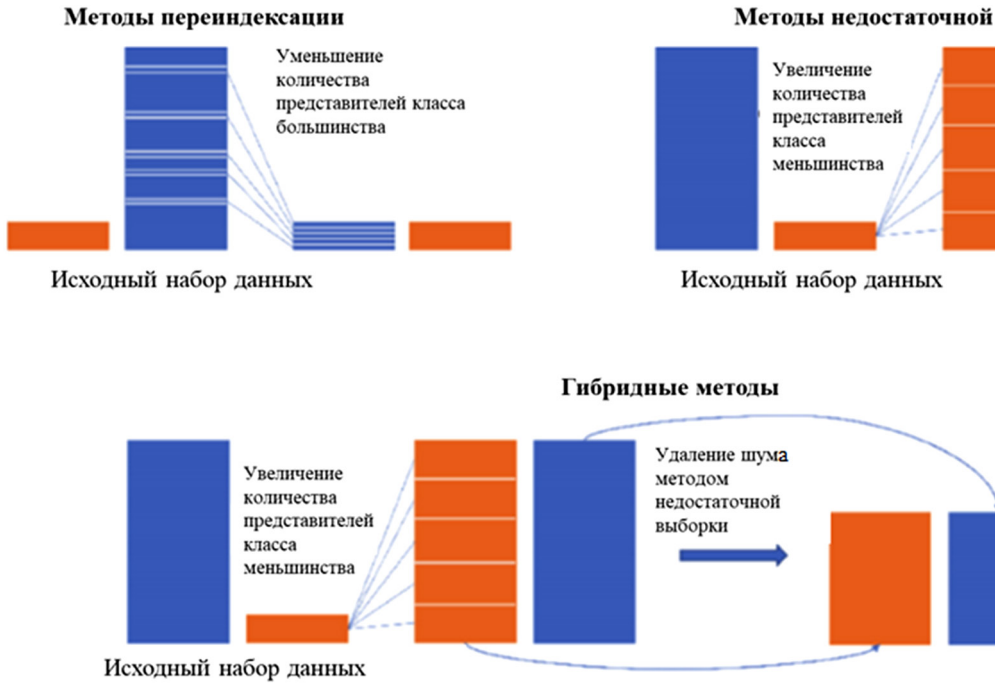


Рис. 5. Стратегии сэмплинга данных

ществует ряд модификаций этого метода, также он стал вдохновением для создания других алгоритмов работы с несбалансированными данными. *SMOTE* возглавляет группу методов сэмпелирования данных – т.е. увеличения количества представителей класса меньшинства. Его суть показана на рис. 4.

Синтетические экземпляры генерируются в «пространстве функций», а не в «пространстве данных». Выборки класса меньшинства пополняются путем ввода синтетических примеров вдоль отрезков линии, соединяющих любого/всех ближайших соседей класса меньшинства *k* [5]. *SMOTE* относится к группе методов оверсэмплинга.

Другое направление работы с несбалансированными данными – уменьшение количества представителей класса большинства (андерсэмплинг или методы переиндексации).

Сочетание различных стратегий сэмплинга – гибридные методы, предусматривающие последовательное применение алгоритмов оверсэмплинга и андерсэмплинга. Визуализация процессов и результатов вышеупомянутых стратегий приведена на рис. 5.

В работе [6] проведен обзор методов работы с дисбалансами классов, информация о них, расширенная современными методами, приведена в табл. 2.

Исходя из преимуществ и недостатков методов, приведенных в табл. 2, а также опыта исследователей, изучающих методы сэмплирования данных [51, 52, 53, 54], в качестве основного метода повышения качества классификации несбалансированных данных выбран метод *NearMiss* [47, 55]. Цель применения метода *NearMiss* – сбалансировать распределение наблюдений по классам меньшинства и большинства на основе оценки расстояния между экземплярами из разных классов.

Табл. 2. Обзор методов работы с дисбалансом классов

Решение	Суть группы методов	Сила	Слабость
Подход на уровне данных			
MLSMOTE [7]	Предварительный обучению этап: используется либо недостаточная выборка, либо избыточная выборка, чтобы уменьшить коэффициент дисбаланса в обучающих данных	Прямой подход и широкое распространение использования	Риск переобучения
Диверсификация недостаточной выборки на основе стохастической чувствительности (*DSUS*) [8]			
ACO-отбор колоний муравьев [9]			
SMOTE [10]			
Эволюция недостаточной выборки [11]			
Бустинг стоимости (важности)			
Чувствительное к стоимости лингвистическое нечеткое правило [12]	Статьи затрат (обозначают важность идентификации классов) обозначают неравномерную важность идентификации среди классов. Повышение стратегии может намеренно смещать обучение в сторону классов, связанных с более высокой важностью идентификации и, в конечном итоге, улучшить производительность идентификации на них	Прямолинейный метод, особенно если известна цена ошибки	Дополнительные затраты на обучение в связи с поиском эффективной матрицы затрат, особенно когда реальная цена ошибки неизвестна
Повышение чувствительности к стоимости [13]			
Отбор признаков			
Выбор характеристик класса меньшинства	Методы отбора признаков, на которых проводится обучение	Помогает решить проблему перекрытия классов	Дополнительные вычислительные затраты из-за необходимости решать задачи предварительной обработки данных
Выбор объекта на основе плотности			
FAST генерация наблюдений на основе *roc*			
CFS генерация наблюдений на основе корреляций			
Подход на уровне алгоритмов			
Argument-based rule learning [14]	Специализированные алгоритмы, которые непосредственно изучают распределение дисбаланса классов из наборах данных	Эффективность за счет модифицированных алгоритмов для обучения исключительно из распределения классов дисбаланса	Может потребоваться решение задач предварительной обработки, чтобы сбалансировать неравномерное распределение классов
Обучение на основе различий [15]			
Нечеткий классификатор [16]			
z-SVM [17]			
Иерархическое нечеткое правило [18]			
Распределение условного ближайшего соседа класса [19]			
Примерное обобщение по *k-NN* [20]			
Взвешенный классификатор ближайшего соседа [21]			
Одноклассовое обучение			
Одноклассовое обучение [22]	Моделирование классификатора на представлении класса меньшинства	Простота применения	Неэффективен при применении с алгоритмами классификации, которые должны учиться на преобладающем классе
Распределение условного ближайшего соседа класса (*CCNND*) [19]			
Экономичное обучение			
Байесовский классификатор *SVM* [23]	Группа методов классификации, основанная на стоимости ошибки классификации как ложного срабатывания, так и промаха	Простой и быстрый способ обработки	Неэффективен, если реальная стоимость недоступна. Вводятся дополнительные расходы, если требуется исследование стоимости, когда цена ошибки неизвестна
Чувствительное к стоимости обучение с *SVM* [24]			
Чувствительный к стоимости *NN* с *PSO* [25]			
SVM для адаптивно асимметричной ошибочной классификации стоимости [26]			

Решение	Суть группы методов	Сила	Слабость
Метод ансамблей			
SMOTE и ансамбль выбора функций [27]	Группа методов, основанных на применении нескольких классификаторов, обучающихся напрямую на данных, и суммирования их оценок для выработки окончательного решения по классификации	Разносторонние подходы	Сложность возрастает с использованием большего количества классификаторов. Трудно достичь концепции разнообразия
Ансамбли GA [28]			
Ансамбли для финансовых проблем [29]			
Бустинг в ансамбле *SVM*-методов [30]			
RUSBoost [31]			
SMOTEBoost [32]			
Гибридный подход			
FTM-SVM	Используется более одного алгоритма машинного обучения для улучшения качества классификации, часто за счет гибридизации с другими алгоритмами обучения для достижения лучших результатов. Гибридизация применяется с целью облегчить проблему выборки, выбора подмножества признаков, оптимизации матрицы затрат и точной настройки классических алгоритмов обучения	Набирает популярность в классификации классового дисбаланса Обучение симбиозу посредством комбинации с другими алгоритмами обучения	Требуется тщательная оценка проекта, чтобы учесть различия между применяемыми методами
Обучение на основе *F*-меры [33]			
Лингвистическое нечеткое правило [34]			
Электронный алгоритм нечеткого классификатора [35]			
Извлечение нечетких правил с помощью *GA* [36]			
Нейро-нечеткость [37]			
Медицинские данные нейросети [38]			
Нейронные сети с *SMOTE* [39]			
Классификатор *kNN* для медицинских данных [40]			
NN обучена с *BP* и *PSO* для медицинских данных [41]			
Ядра дерева зависимостей [42]			
Использование чувствительности к стоимости в дереве [43]			
Недодискретизация и *GA* для *SVM* [44]			
Другие методы			
ADASYN [45, 46]	После создания выборки он добавляет случайные небольшие значения к точкам. Другими словами, вместо того, чтобы вся выборка линейно коррелировала с родителем, в них немного больше различий	Адаптивный характер: количество синтетических наблюдений основывается на соотношении наблюдений большинства и меньшинства. *ADASYN* делает больший акцент на более сложных областях данных	Недостатком *ADASYN* является то, что на него легко влияют выбросы
NearMiss [45, 47]	Случайное исключение примеров классов большинства. Когда экземпляры двух разных классов находятся очень близко друг к другу, мы удаляем экземпляры класса большинства, чтобы увеличить пространство между двумя классами. Это помогает в процессе классификации	Может уменьшить перекрытие выборок между различными классами	При недостаточной выборке количество недостаточных выборок невозможно контролировать, а массовые выборки, которые можно исключить, ограничены. Лучше всего использовать в качестве метода очистки данных в сочетании с другими методами

Решение	Суть группы методов	Сила	Слабость
Edited Nearest Neighbours [45, 48]	Представители класса большинства исключаются, если более половины из *K* их соседей не принадлежат к классу большинства	Уменьшает ошибку неверной классификации представителя класса большинства	Не может контролировать количество недостаточной выборки
Tomek Links [3, 45, 49]	Исключаются представители класса большинства, находящиеся в непосредственной близости от объектов класса меньшинства	Все образцы, которые являются ближайшими соседями друг к другу, принадлежат одному и тому же классу, который может быть лучше классифицирован	
Neighbourhood Cleaning Rule [45]	Эта стратегия также направлена на то, чтобы удалить те примеры, которые негативно влияют на исход классификации классов меньшинства. Для этого все примеры классифицируются по правилу трех ближайших соседей	Повышает точность классификации представителей классов меньшинства	Может привести к такому уменьшению объема выборки, которого будет недостаточно для обучения
Condensed Nearest Neighbour [45, 50]	Метод позволяет находить отличие между похожими примерами, но принадлежащими к разным классам.	Поиск отличий между похожими примерами, но принадлежащими к разным классам.	

Алгоритм реализации *NearMiss* библиотеки *imblearn Python* (библиотека включает в себя набор инструментов для несбалансированного набора данных в машинном обучении) предусматривает 3 стратегии отбора представителей классов, а именно:

1 стратегия (*version*=1). Выбор наблюдений из класса большинства, для которых среднее расстояние до *k* ближайших наблюдений из класса меньшинства является наименьшим (по умолчанию *k*=3, вариативный параметр при обучении). Будут сохранены только те наблюдения без опасных событий, которые наиболее близки к наблюдениям с опасными событиями;

2 стратегия (*version*=2). Выбор наблюдений из класса большинства, для которых среднее расстояние до *k* самых дальних наблюдений класса меньшинства является наименьшим (по умолчанию *k*=3, вариативный параметр при обучении). Будут сохранены только те наблюдения без опасных событий, которые находятся в центре масс пересечения множеств классов большинства и меньшинства;

3 стратегия (*version*=3). Сначала для каждого наблюдения из класса большинства будут сохранены их *M* ближайших соседей (по умолчанию *M*=3, вариативный параметр при обучении). Затем выбираются наблюдения из класса меньшинства, для которых среднее расстояние до *N* ближайших соседей является наибольшим (по умолчанию *N*=3, вариативный параметр при обучении).

В отличие от исследований [1–4, 52, 53], объектом нашего внимания является не показатель точности или ошибки классификации, а детализация точности прогноза опасных событий и точности прогноза появления инцидентов. Кроме этого, из-за того, что 98% наблюдений относятся к классу большинства, следует ожидать, что точность большинства обученных классификаторов будет составлять 98%. В таких условиях легко прийти к выводам некоторых исследователей о неэффективности методов балансировки классов.

В ходе исследования использовалось сочетание нескольких методов повышения качества классификации несбалансированных данных, а именно сочетание методов подбора параметров моделей (*GridSearchCV*) с методом *NearMiss* и различные функции качества.

Варьируемые в ходе эксперимента параметры:

1) в рамках функции *GridSearchCV:*

(1) *GBS: random_state, tol, max_depth*;

(2) *Logi: tol, class_weight, max_iter, solver, random_state, C*;

(3) *KNN: n_neighbors, weights, metric*;

(4) *DTC: criterion, max_depth, min_samples_split, max_features*;

2) функции качества: *max_error, balanced_accuracy,accuracy, neg_log_loss, explained_variance, neg_mean_squared_error, neg_mean_squared_log_error, neg_median_absolute_error, r2*;

3) параметры *NearMiss*: стратегии отбора представителей классов, количество представителей класса меньшинства, количество представителей класса большинства.

Схема эксперимента:

1) разделение выборки на основную и валидационную. К основной выборке относятся наблюдения, полученные до 2019 года, к валидационной – после 2019 года;

2) разделение основной выборки на тренировочную и тестовую. Метод – *train_test_split*, размер тестовой выборки – 20 %.

В ходе экспериментов определено, что лучшие показатели точности достигаются при функции качестве *neg_log_loss*. В виду чрезвычайно малого количества представителей класса меньшинства (менее 2%) качество варьирования этого показателя не оказывает влияния на качество классификации. В ходе большинства экспериментов количество представителей класса меньшинства было фиксированным и равно максимально возможному.

Рассмотрим результаты экспериментов. На рис. 6 приведен график зависимости точности прогноза методом *GBC* появления событий в зависимости от количества представителей класса большинства на основной выборке со стратегий *NearMiss* 2.

Рис. 6. Точность классификации на основной выборке, модель *GBC*, *NearMiss* 2

Стартовые характеристики точности (без применения метода *NearMiss*): точность на тестовой части обучающей выборки: 0,983; точность прогноза опасного события: 0,416; точность прогноза появления неопасного инцидента: 1. Как видно из рис. 6, существует точка пересечения графиков точности классификации опасных событий и инцидентов. По сути, чем выше будет находиться эта точка, тем точнее будет итоговая модель. Аналогичные графики получены для стратегии сэмплирования *NearMiss* 1. Точность в 70% выглядит хорошим результатом при стартовой точности классификации опасного события 41,6%. Следующий эксперимент был проведен при стратегии отбора № 2. Для удобства восприятия, на рис. 7 приведена только область пересечения графиков, характеризующих валидационную выборку.

Рис. 7. Точность на валидационной выборке, модель *GBC*, *NearMiss* 1.

Точность работы классификатора на валидационной выборке (в-выборке) оказалась существенно выше точности на основной выборке. Анализ показал, что это связанно с объемом в-выборки (классификатор не «успел» наделать ошибок), а также месяцами, для которых делался прогноз. Точность за аналогичные периоды из в-выборки в основной выборке оказалась выше, чем в среднем по всей выборке.

Прежде, чем перейти к анализу других данных, вернемся к вопросу использования показателя точности классификатора как характеристики эффективности. Точность классификатора *GBC* на несбалансированных данных составляла 0,983. При этом точность прогноза опасного события на в-выборке – 0. На сбалансированных данных точность составила 0,9811 (ниже, чем на оригинальном объеме выборки), а точность прогноза опасного события на в-выборке – 1,0, точность прогноза появления инцидента – 0,934. Аналогичные показатели были получены для других классификаторов. Таким образом, проводя исследования точности классификации несбалансированных данных, нельзя опираться на один «свернутый» показатель качества.

На рис. 8 и 9 приведены графики точности модели *Logi* для основной выборки и в-выборки на стратегии *NearMiss* 1.

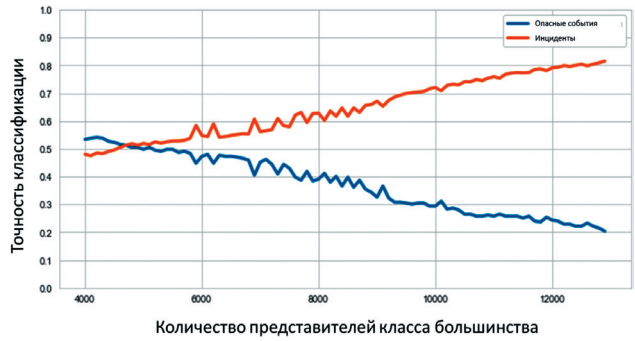


Рис. 8. Точность на основной выборке, модель *Logi*, *NearMiss* 1

Рис. 9. Точность на валидационной выборке, модель *Logi*, *NearMiss* 1

Из графиков на рис. 8 и 9 видно, что лучшие показатели точности для основной и в-выборки достигаются при различных балансах классов. Это свидетельствует о том, что отвергнутые на основании данных основной

выборки модели могут давать лучшие результаты на новых данных. Один из методов решения указанной проблемы – применение ансамблевых методов [56]. Основная идея изучения ансамбля классификаторов состоит в том, чтобы построить несколько классификаторов из исходных данных, а затем агрегировать свои прогнозы при классификации неизвестных образцов (рис. 10). Применение таких методов является перспективой исследования.

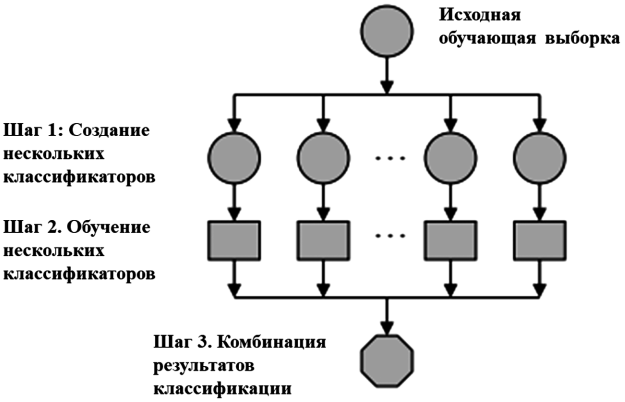


Рис. 10. Общий алгоритм работы ансамблевых методов

Анализ и выводы

По мере развития систем безопасности количество опасных событий становится все меньше, но при этом растет цена потенциальных последствий. Среди инцидентов с ЖЭс доля опасных событий не превышает 2% в год. Для повышения точности классификации и прогнозирования типов событий в условиях дисбаланса класса необходимо принять методы балансирования выборки.

В настоящее время активно используется более 50 методов, позволяющих работать с несбалансированными выборками. Однако в известных авторам публикациях не рассматривается вопрос анализа совместного изменения точности прогнозирования представителей классов меньшинства и большинства. В статье представлены графики точностей классификации инцидентов с объектами ЖЭс на обучающей и валидационной выборке.

Предложена схема работы с несбалансированными данными, включающая в себя сочетание нескольких методов повышения качества классификации несбалансированных данных, а именно: настройка параметров моделей, выбор показателя качества и соотношения количества представителей классов меньшинства и большинства с помощью метода *NearMiss*.

Применение методов работы с дисбалансом классов позволяет существенно повысить точность прогнозирования появления представителя класса меньшинства (опасного события) с 0 до 70-90%. Однако при повышении точности прогнозирования редкого события понижается точность прогнозирования появления представителя класса меньшинства.

Перспективами исследования являются: применение гибридных методов классификации и прогнозирования событий, а также разработка метрики качества обучения, основанной на характеристике точки пересечения графиков точности классификации представителей различных классов.

Библиографический список

1. Севастьянов Л.А, Шетиние Е.Ю. О методах повышения точности многоклассовой классификации на несбалансированных данных // Информатика и ее применение. 2020. Том 14. № 1. С. 63-70.

2. Садов М.А. Исследование методов классификации текстов для несбалансированных данных // Полиматис. 2016. № 2. С. 28-41.

3. Маслихов С.Р., Мохов А.С., Толчеев В.Ю. Построение сбалансированных классов в задаче классификации запросов пользователей // Сборник трудов V Международной научно-практической конференции «Дистанционные образовательные технологии». 2020. С. 245-248.

4. Шипицын А.В., Журавлева Н.В. Оценка онлайн-заявок на ипотечный кредит с помощью алгоритмов Machine Learnng // Вестник Белгородского университета кооперации, экономики и права. 2016. № 4(60). С. 199-209.

5. Chawla N.V., Bowyer W.B., Hall L.O. et al. SMOTE: Synthetic Minority Over-sampling Technique // Journal of Artificial Intelligence Research. 2002. № 16. P. 321-357.

6. Ali A., Shamsuddin S.M., Ralescu A. Classification with class imbalance problem: a review // International Journal of Advances in Soft Computing. 2013. № 7. P. 176-204.

7. Mladenic D., Grobelnik M. Feature selection for unbalanced class distribution and naive bayes // Proceedings of the Sixteenth International Conference on Machine Learning (ICML 1999), Bled, Slovenia, June 27–30, 1999. P. 258–267.

8. Yang T.-N., Wang S.-D. Robust algorithms for principal component analysis // Pattern Recognition Letters. 1999. 20(9). P. 927-933.

9. Yu H., Ni J., Zhao J. ACOSampling: An ant colony optimization-based undersampling method for classifying imbalanced DNA microarray data // Neurocomputing. 2013. 101(0). P. 309-318.

10. Chawla N.V. SMOTE: synthetic minority oversampling technique // arXiv:1106.1813, 2002.

11. García S., Herrera F. Evolutionary undersampling for classification with imbalanced datasets: Proposals and taxonomy // Evolutionary Computation. 2009. 17(3). P. 275-306.

12. Yin L. Feature selection for high-dimensional imbalanced data // Neurocomputing. 2013. 105(0). P. 3-11.

13. Sun Y. Cost-sensitive boosting for classification of imbalanced data // Pattern recognition. 2007. 40(12). P. 3358-3378.

14. Luukka P. Nonlinear fuzzy robust PCA algorithms and similarity classifier in bankruptcy analysis // Expert Systems with Applications. 2010. 37(12). P. 8296-8302.

15. Zheng Z., Wu X., Srihari R. Feature selection for text categorization on imbalanced data // ACM SIGKDD Explorations Newsletter. 2004. 6(1). P. 80-89.

16. Visa S., Ralescu A.L. Fuzzy Classifiers for Imbalanced Data Sets. University of Cincinnati, Computer Science Dept. Cincinnati, OH, United States, 2007. 157 p.

17. Imam T., Ting K., Kamruzzaman J. z-SVM: An SVM for Improved Classification of Imbalanced Data // AI 2006: Advances in Artificial Intelligence. 19th Australian Joint Conference on Artificial Intelligence, Hobart, Australia, December 4-8, 2006. Proceedings. P. 264-273.

18. Fernández A., M.J. del Jesus, Herrera F. Hierarchical fuzzy rule based classification systems with genetic rule selection for imbalanced data-sets // International Journal of Approximate Reasoning. 2009. 50(3). P.561-577.

19. Kriminger E., Principe J.C., Lakshminarayan C. Nearest Neighbor Distributions for imbalanced classification // The 2012 international joint conference on neural networks (IJCNN), Brisbane, QLD, Australia, 10–15 June 2012. P. 1–5.

20. Li Y., Zhang X. Improving k nearest neighbor with exemplar generalization for imbalanced classification // Advances in knowledge discovery and data mining: 15th Pacific-Asia Conference, PAKDD 2011, Shenzhen, China, May 24-27, 2011, Proceedings, Part II. P. 321-332.

21. Candès E.J. Robust principal component analysis // Journal of the ACM (JACM). 2011. 58(3). P. 11.

22. Japkowicz N., Myers C., Gluck M. A novelty detection approach to classification // IJCAI. Vol. 1. 1995. P. 518–523.

23. Jolliffe I. Principal component analysis // Encyclopedia of Statistics in Behavioral Science. John Wiley & Sons, Ltd, 2005.

24. Cao P., Zhao D., Zaiane O. An Optimized Cost-Sensitive SVM for Imbalanced Data Learning // Advances in Knowledge Discovery and Data Mining: 17th Pacific-Asia Conference, PAKDD 2013, Gold Coast, Australia, April 14-17, 2013, Proceedings, Part II. P. 280-292.

25. Cao P., Zhao D., Zaiane O. A PSO-based Cost-Sensitive Neural Network for Imbalanced Data Classification // Trends and Applications in Knowledge Discovery and Data Mining: PAKDD 2013 International Workshops: DMApps, DANTH, QIMIE, BDM, CDA, CloudSD, Gold Coast, QLD, Australia, April 14-17, 2013, Revised Selected Papers. P 452–463.

26. Wang X., Shao H, Japkowicz N et al. Using SVM with Adaptively Asymmetric Misclassification Costs for Mine-Like Objects Detection // 2012 11th International Conference on Machine Learning and Applications, Boca Raton, FL, USA, 12-15 Dec. 2012. P. 78–82.

27. Yang P., Liu W, Zhou B.B. et al. Ensemble-based wrapper methods for feature selection and class imbalance learning // Advances in Knowledge Discovery and Data Mining: 17th Pacific-Asia Conference, PAKDD 2013, Gold Coast, Australia, April 14-17, 2013, Proceedings, Part I. P. 544-555.

28. Yu E., Cho S. Ensemble based on GA wrapper feature selection // Computers & Industrial Engineering. 2006. 51(1). P.111-116.

29. Liao J.-J. An ensemble-based model for two-class imbalanced financial problem // Economic Modelling. 2014. 37(0). P.175-183.

30. Liu Y., An A., Huang X. Boosting prediction accuracy on imbalanced datasets with SVM ensembles // Advances in Knowledge Discovery and Data Mining, 10th Pacific-Asia Conference, PAKDD 2006, Singapore, April 9-12, 2006, Proceedings. P. 107-118.

31. Seiffert C. RUSBoost: A hybrid approach to alleviating class imbalance // Systems, Man and Cybernetics. Part A: Systems and Humans. IEEE Transactions. 2010. 40(1). P.185-197.

32. Chawla N.V. SMOTEBoost: Improving prediction of the minority class in boosting, // Proceedings of Conference: Knowledge Discovery in Databases: PKDD 2003, 7th European Conference on Principles and Practice of Knowledge Discovery in Databases, Cavtat-Dubrovnik, Croatia, September 22-26, 2003. P. 107-119.

33. Wasikowski M., Chen X.-W. Combating the small sample class imbalance problem using feature selection // Knowledge and Data Engineering. IEEE Transactions. 2010. 22(10). P.1388-1400.

34. Martino M.D. Novel Classifier Scheme for Unbalance Problems // Pattern Recognition Letters. 2013. Vol. 34. Issue 10. P. 1146–1151.

35. Fernández A. A study of the behaviour of linguistic fuzzy rule based classification systems in the framework of imbalanced data-sets // Fuzzy Sets and Systems. 2008. 159(18). P. 2378-2398.

36. Le X., Mo-Yuen C., Taylor L.S. Power Distribution Fault Cause Identification With Imbalanced Data Using the Data Mining-Based Fuzzy Classification E-Algorithm // Power Systems. IEEE Transactions. 2007. 22(1). P. 164-171.

37. Soler V. Imbalanced Datasets Classification by Fuzzy Rule Extraction and Genetic Algorithms // Data Mining Workshops 2006. ICDM Workshops 2006. Sixth IEEE International Conference 2006.

38. Hung C.-M. Huang Y.-M. Conflict-sensitivity contexture learning algorithm for mining interesting patterns using neuro-fuzzy network with decision rules // Expert Systems with Applications. 2008. 34(1). P. 159-172.

39. Jeatrakul P., Wong K.W., Fung C.C. Classification of imbalanced data by combining the complementary neural network and SMOTE algorithm // Proceedings of the 17th international conference on Neural information processing: models and applications. Volume Part II. 2010. Springer-Verlag: Sydney, Australia. P. 152-159.

40. Malof J.M., Mazurowski M.A., Tourassi G.D. The effect of class imbalance on case selection for case-based classifiers: An empirical study in the context of medical decision support // Neural Networks. 2012. 25(0). P. 141-145.

41. Mazurowski M.A. Training neural network classifiers for medical decision making: the effects of imbalanced datasets on classification performance // Neural networks. 2008. 21(2-3). P. 427-436.

42. Culotta A. Sorensen J. Dependency tree kernels for relation extraction // Proceedings of the 42nd Annual Meet-

ing on Association for Computational Linguistics. Association for Computational Linguistics, 2004.

43. Drummond C., Holte R.C. Exploiting the cost (in) sensitivity of decision tree splitting criteria // ICML. 2000.

44. Al-Shahib A., Breitling R., Gilbert D. Feature selection and the class imbalance problem in predicting protein function from sequence // Applied Bioinformatics. 2005. 4(3). P. 195-203.

45. Koziarski M. Radial-Based Undersampling for imbalanced data classification // Pattern Recognition. 2020. Vol. 102.

46. He H., Bai Y., Garcia E.A. et al. ADASYN: adaptive synthetic sampling approach for imbalanced learning // IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence). 2008. P. 1322-1328.

47. Mani I., Zhang I. kNN approach to unbalanced data distributions: a case study involving information extraction // Proceedings of Workshop on Learning from Imbalanced Datasets. 2003. Vol. 126.

48. Wilson D.L. Asymptotic properties of nearest neighbor rules using edited data // IEEE Trans. Syst. Man Cybern. 1972. 2 (3). P. 408-421.

49. Tomek I. Two modifications of CNN // IEEE Trans. Syst. Man Cybern. 1976. 6. P. 769-772.

50. Hart P. The condensed nearest neighbor rule // IEEE Trans. Inf. Theory. 1968. 14(3). P. 515-516.

51. Махсотова Ц.В. Исследование методов классификации при несбалансированности классов // Научный журнал. 2017. №5 (18). URL: https://cyberleninka.ru/article/n/issledovanie-metodov-klassifikatsii-pri-nesbalansirovannosti-klassov (дата обращения: 5 июля 2020 г.).

52. Каврин Д.А., Субботин С.А. Методы количественного решения проблемы несбалансированности классов // Радіоелектроніка, інформатика, управління. 2018. №1 (44). URL: https://cyberleninka.ru/article/n/metody-kolichestvennogo-resheniya-problemy-nesbalansirovannosti-klassov (дата обращения: 6 июля 2020 г.).

53. Yi Lu, Hong Guo, Feldkamp L. Robust neural learning from unbalanced data samples // 1998 IEEE International Joint Conference on Neural Networks Proceedings. IEEE World Congress on Computational Intelligence (Cat. No.98CH36227), Anchorage, AK. 1998. Vol. 3. P. 1816-1821.

54. Al-Stouhi S., Reddy C.K. Transfer learning for class imbalance problems with inadequate data // Knowledge and Information Systems. 2016. 48. P. 201–228.

55. Near-Miss – version 0.9.0.dev0. API reference. URL: https://imbalanced-learn.org/dev/references/generated/imblearn.under_sampling.NearMiss.html (дата обращения: 10 июля 2021 г.).

56. Sun Y. Cost-Sensitive Boosting for Classification of Imbalanced Data // Pattern Recognition. 2007. Vol. 40. Issue 12. P. 3358–3378.

Сведения об авторах

Ольга Борисовна Проневич – начальник отдела АО «НИИАС», ул. Нижегородская, д. 27, стр. 1, Москва, Российская Федерация, 109029, тел +7 (495) 786-68-57; e-mail: obpronevich@gmail.com

Михаил Вадимович Зайцев – ведущий специалист АО «НИИАС», ул. Нижегородская, д. 27, стр. 1, Москва, Российская Федерация, 109029, тел +7 (495) 786-68-57; e-mail: m.v.zaicev@mail.ru

Вклад авторов

Проневич О.Б. исследовала проблему дисбаланса классов при прогнозировании появления опасного события на объектах железнодорожного электроснабжения, провела анализ существующих методов работы с несбалансированными данными, а также серию экспериментов по исследованию влияния различных соотношений количества представителей классов меньшинства и большинства. Также автором определены перспективы исследования.

Зайцев М.В. провел анализ популярных методов классификации событий в условиях дисбаланса классов, провел исследование влияния различных стратегий отбора *NearMiss* на качество классификации опасных событий на железнодорожном транспорте.

Конфликт интересов

Авторы заявляют об отсутствии конфликта интересов.