pISSN 2071-4688
eISSN 2311-8709

Банковская деятельность

ПРЕОДОЛЕНИЕ НЕСБАЛАНСИРОВАННОСТИ КЛАССОВ ПРИ МОДЕЛИРОВАНИИ ДЕФОЛТА КРЕДИТНОГО ТРЕБОВАНИЯ

Владислав Владимирович РОСКОШЕНКО

аспирант, магистр экономики, экономический факультет, МГУ имени М.В. Ломоносова,
Москва, Российская Федерация
roskoshenkoeco@mail.ru
ORCID: отсутствует
SPIN-код: отсутствует

История статьи:
Рег. № 667/2019
Получена 17.10.2019
Получена в доработанном виде 31.10.2019
Одобрена 14.11.2019
Доступна онлайн 29.11.2019

УДК 336.71
JEL: G21, G28

Ключевые слова:
кредитный скоринг, логистическая регрессия, ансамбли, несбалансированность классов, бинарная классификация

Аннотация

Предмет. Проблема несбалансированности классов в выборочных данных при моделировании дефолта кредитного требования, подходы к предварительной обработке данных, позволяющие преодолеть дисбаланс классов. Имеющиеся исследования по сопоставлению таких подходов выполнены либо в отношении небольшого числа методов, либо на специфических данных из отдельных областей деятельности. Ранее в литературе не рассмотрены подходы на основе сочетания методов предварительной обработки данных с ансамблевым решением (стэкингом).

Цели. Произвести поиск оптимального варианта по преодолению проблемы несбалансированности классов среди каждой из групп подходов для банковских данных о кредитовании физических лиц.

Методология. Использованы математическое моделирование, статистический анализ и контент-анализ источников.

Результаты. Показано, что подход EditedNearestNeighbours, будучи довольно сложным с вычислительной точки зрения, оказался оптимальным. В его основе — удаление представителей доминирующего класса, плохо удовлетворяющих своему окружению, которое определяется посредством кластеризации. Среди сочетаний подходов предварительной обработки данных и стэкинга оптимальным оказался вариант с RandomOverSampler. Последний предполагает увеличение доли миноритарного класса случайным образом и является одним из наиболее простых.

Область применения. Результаты могут быть использованы в кредитном скоринге и в любом статистическом моделировании, где требуется бинарная классификация.

Выводы. Осуществлено исчерпывающее сопоставление подходов по преодолению проблемы несбалансированности классов в выборочных данных. Были определены оптимальный подход среди подходов предварительной обработки данных, а также оптимальное сочетание подхода предварительной обработки данных с ансамблевым решением.

Для цитирования: Роскошенко В.В. Преодоление несбалансированности классов при моделировании дефолта кредитного требования // Финансы и кредит. — 2019. — Т. 25, № 11. — С. 2534 — 2561.
https://doi.org/10.24891/fc.25.11.2534

Введение

Проблема несбалансированности классов в выборочных данных имеет широкое распространение в задачах классификации. Так, неравномерное распределение классов можно встретить при моделировании постановки медицинских диагнозов, распознавании лиц, сортировке писем, поиске аномалий. Очевидно, данная проблема характерна для моделирования дефолта кредитных требований в кредитовании физических лиц.

Классификация на основе несбалансированных данных заведомо имеет существенный недостаток эффективности при использовании любых стандартных алгоритмов. Дело в том, что последние ожидают относительно сбалансированные данные и равные издержки

неправильной классификации между классами, отмечает Й. Сан [1].

Перекошенное (неравномерное) распределение классов в данных обычно приводит к «переобученным» моделям, чья обобщающая способность ниже требуемой. Также данная проблема имеет, как правило, ряд других проблем, среди которых проблемы наслаивания классов, малого размера выборки, а также разобщенности.

Проблема наслаивания классов возникает, когда часть множества представителей меньшего класса накладывается на множество другого класса в пространстве доступных характеристик. Очевидно, в таких ситуациях классификатор будет склоняться к обобщающим правилам, относящим все наблюдения данного региона пространства к доминирующему классу. В. Гарсия и др. отмечают, что присутствие проблемы наслаивания может сильно осложнить классификацию при несбалансированности классов. Более того, осложнить классификацию может отличие в соотношении классов в целом по данным и локально в области наслаивания [2].

Проблема малого размера выборки — недостаточное количество представителей миноритарного класса для адекватного моделирования. С. Стефен и др. отмечают, что при фиксировании несбалансированного соотношения классов, количественное снижение меньшего класса приводит к росту уровня ошибок классификации [3].

Проблема разобщенности имеет место, если представители класса сгруппированы отдельными сегментами в пространстве доступных характеристик. В таком случае довольно сложным оказывается формулирование общей концепции конкретного класса. Очевидно, что наличие множества малочисленных групп представителей для моделирования оказывается более сложной задачей, нежели наличие меньшего количества более крупных объединений объектов. Соответственно, Г.М. Вайс и др. заключают, что проблема разобщенности именно в отношении представителей меньшего класса ведет к большим ошибкам классификации в силу недостаточного их количества в отдельных группах [4].

Другим негативным последствием несбалансированности классов в выборочных данных является ложная оптимизация моделирования посредством стандартных критериев качества модели. Так, если миноритарный класс составляет 1%, стандартные критерии, например точность, будут склонять процесс моделирования к формированию классификатора, игнорирующего наличие малочисленного класса и просто прогнозирующего отсутствие такового (с уровнем точности, соответственно, 99%).

Для решения проблемы несбалансированности данных в научной литературе было предложено множество подходов. Эти подходы условно можно разделить на группы в зависимости от того, как решается проблема (*рис. 1*).

Подходы на уровне алгоритмов требуют такой модификации алгоритмов обучения, чтобы они лучше справлялись с нахождением представителей миноритарного класса. Для использования этого подхода необходимо глубокое понимание природы классификатора и области применения полученных моделей. В рамках данного направления возможно отметить ряд работ.

Ученые Й. Лин и др. предложили адаптацию алгоритма Машины Опорных Векторов (SVM) для нестандартных случаев. Ими могут быть разные издержки неправильной классификации, а также разное распределение классов внутри целевой выборки (например, будущие поколения клиентов в кредитном скоринге) и выборочной (например, исторические выдачи ссуд) [5]. Последнее может быть результатом либо дрейфа популяции (изменением характеристик поколений) во времени, либо намеренным отбором классов в определенном соотношении, не совпадающим с изначальным. Также для SVM Г. Ву и др. предложили адаптацию к условиям несбалансированности посредством изменения матрицы ядра [6].

Подходы на уровне данных решают проблему несбалансированности на этапе подготовки данных для моделирования. Таким образом, отсутствует зависимость от конкретного классификатора. Имеются два основных направления: расширение доли миноритарного класса и снижение численности доминирующего класса. В первую очередь стоит выделить подходы по увеличению присутствия меньшего класса (*рис. 2*).

Исследователи Н.В. Чавла и др. предложили так называемый синтетический подход (the Synthetic Minority Oversampling Technique, SMOTE) к увеличению количества представителей миноритарного класса [7]. В основе подхода — использование метода *k*-ближайших соседей. Для случайно выбранного представителя малого класса определяются *k*-ближайших соседа. Случайно выбранный из этих соседей и изначальный представители будут использованы для интерполяции — взвешенного усреднения их характеристик. Интерполированные характеристики лягут в основу нового «синтетического» представителя.

Ученые Х. Хи и др. разработали адаптивный синтетический подход (the Adaptive Synthetic, ADASYN), во многом похожий на SMOTE [8]. Отличие заключается в прямой зависимости количества созданных новых представителей от свойств изначального. Так, большее количество «синтетических» представителей создается из окружения, если в окрестность соседей попадает большее количество представителей противоположного (доминирующего) класса. Иными словами, чем сложнее для кластеризации методом *k*-ближайших соседей верно определить представителя меньшего класса, тем больше на его основе будет создано «синтетических» представителей. В данном смещении фокуса классификации в сторону «трудных» примеров и проявляется адаптивность подхода.

Оба рассмотренных подхода могут давать субоптимальные результаты, так как SMOTE может интерполировать «выбросы» и регулярные точки, а ADASYN стремится полагаться на «выбросы». Для устранения этой возможной неоптимальности были предложены модификации SMOTE.

Исследователи Х. Хэн и др. предложили BorderlineSMOTE, который предварительно группирует представителей миноритарного класса на «выбросы» (все соседи представляют противоположный класс), «находящиеся в опасности» (хотя бы половина соседей представляет собственный класс) и «безопасные» (все представители собственного класса) [9]. Именно вторая группа используется для синтезирования новых представителей. Разновидность данного подхода — Borderline-1 SMOTE для интерполяции выбирает представителей меньшего класса из окружения, а Borderline-2 SMOTE — любых.

Ученые Э. Купер и др. использовали Машину опорных векторов для разделения классов (SVMSMOTE) [10]. Элементы выборки, лежащие вдоль разделяющей гиперплоскости (опорные вектора), используются для синтезирования.

Исследователи Ф. Ласт и др. разработали KMeansSMOTE, который предполагает предварительное использование кластеризации методом *k*-средних и последующее приложение SMOTE [11]. Предварительная кластеризация объединяет представителей, и синтезирование начинает зависеть от плотности кластера.

Кроме того, выделяются подходы по уменьшению присутствия доминирующего класса (*рис. 3*). Один из них — создание прототипов имеющихся представителей большего класса (ClusterCentroids). В основе метода лежит использование алгоритма *k*-средних. ClusterCentroids объединяет всех представителей доминирующего класса в *N* (задаваемое конечное количество этих представителей) кластеров посредством алгоритма *k*-средних, а координаты центроидов (центров кластеров) и составляют используемые далее прототипы.

Другое направление объединяет подходы по отбору прототипов из исходной выборки, а не их создание. Можно выделить две большие

группы: подходы с контролируемым (с точки зрения задания конечного количества представителей такого класса) снижением доли большего класса и подходы с неконтролируемым снижением.

Среди контролируемых подходов помимо случайного удаления представителей доминирующего класса необходимо отметить подход NearMiss, предложенный И. Цангом и др. [12]. NearMiss использует в своей основе метод ближайшего соседа и предлагает три эвристики выбора представителей для удаления из выборки данных.

Первая версия NearMiss отбирает представителей, для которых среднее расстояние в пространстве признаков до ближайших N представителей противоположного класса минимальное. Таким образом, в выборке остаются сложные для классификации наблюдения, так как они достаточно похожи на представителей меньшего класса.

Вторая эвристика NearMiss поочередно отбирает представителей, для которых среднее расстояние в пространстве признаков до наиболее дальних *N* представителей противоположного класса минимальное. Таким образом, в выборке также сохраняются наблюдения сложные для классификации, поскольку они достаточно близки к множеству представителей меньшего класса.

Третья эвристика NearMiss представляет двушаговый алгоритм. Вначале для каждого представителя миноритарного класса сохраняются *M* ближайших соседей из противоположного класса. Далее отбираются представители, для которых среднее расстояние в пространстве признаков до ближайших *N* представителей противоположного класса максимальное.

Проблема первой эвристики NearMiss в том, что ее эффективность в значительной степени может быть снижена присутствием шума в данных, в первую очередь в данных меньшего класса. Так, отбор будет сосредотачиваться вокруг точек-выбросов, что приведет к неверному отбору прототипов. Остальные эвристики менее чувствительны к шуму в данных.

Подходы второй группы (неконтролируемые снижения) в целом предназначены для очистки выборки данных от излишних элементов и снижения размерности пространства признаков. В силу этого отсутствует возможность задания требуемой доли того или иного класса к концу работы данных алгоритмов.

В первую очередь стоит отметить подход связей И. Томека [13]. Алгоритм ищет в данных пару ближайших соседей из представителей разных классов, в дальнейшем один из них или оба удаляются из выборки.

Отдельная подгруппа подходов сформирована алгоритмами коррекции выборки данных на основе кластеризации методом ближайшего соседа. Так, Д. Вилсон предложил EditedNearestNeighbours, который удаляет элементы выборки, плохо удовлетворяющие своему окружению [14]. Для каждого представителя доминирующего класса определяются ближайшие *N* соседей. Если не выполняется условие, что все они или большинство (в зависимости от версии подхода) принадлежат к тому же классу, то элементы выборки удаляются. И. Томек модифицировал данный подход посредством множественного повторения алгоритма EditedNearestNeighbours и назвал его RepeatedEditedNearestNeighbours [13]. Также он продолжил изменения и AllKNN, который увеличивает количество соседей с каждой итерацией алгоритма EditedNearestNeighbour.

Отдельная группа подходов использует метод 1-ближайшего соседа, за что носит название сжатых ближайших соседей. П. Харт предложил алгоритм CondensedNearestNeighbour, в основе которого метод 1-ближайшего соседа [15]. Все представители большего класса выделяются в отдельное множество и для каждого элемента осуществляется поиск ближайшего соседа. Тех, для кого он не был найден, добавляют в конечную выборку данных. Логика алгоритма в том, чтобы оставить наиболее уникальных представителей доминирующего класса.

Однако, очевидно, подход чувствителен к шуму в данных. М. Кабат и др. модифицировали данный подход — OneSidedSelection, добавив удаление шума посредством поиска связей И. Томека и применив кластеризацию ко всем элементам выборки [16].

Последняя группа подходов по снижению доли доминирующего класса предлагает обучать классификатор на выборочных данных и удалять представителей с наименьшей предсказанной вероятностью. Так, М.Р. Смит и др. предложили InstanceHardnessThreshold [17].

Вернемся к классификации верхнего уровня подходов по преодолению проблемы несбалансированности (*рис. 1*). Подходы, предполагающие адаптацию алгоритма и изменение данных, заключаются в распределении весов элементам выборки. Эти веса означают различные издержки неправильной классификации тех или иных преставителей классов. Таким образом, адаптированный к учету весов алгоритм классификации искусственно склоняется к более точному распознаванию элементов меньшего класса, так как издержки последних наибольшие.

Ученый П. Домингос предложил процедуру формирования чувствительности классификатора к издержкам неверной классификации — MetaCost [18]. Ч. Линг и др. разработали алгоритм деревьев решений, который учитывает издержки неправильной классификации [19].

Последняя группа решений проблемы несбалансированности классов обращается к ансамблевым решениям. Суть ансамблирования состоит в обучении ряда различных между собой классификаторов и соединении их решений. Существуют три основных вида ансамблей (*рис. 4*).

Бустинг был предложен Р. Шапире в 1990 г. [20]. Как следовало из названия работы, бустинг демонстрировал, что объединение усилий «слабых» классификаторов контринтуитивно дает возможность получить достаточно сильный классификатор. Среди семейства бустингов наиболее широкое признание получили AdaBoost и его модификация — AdaBoost.M1, предложенные Р. Шапире и Й. Фрондом [21], а также более поздняя модификация от Р. Шапире и Й. Сингера — AdaBoost.M2 [22]. AdaBoost обучает классификатор (дерево решений) на всех данных для обучения. Однако с каждой итерацией начинает уделять больше внимания неправильно классифицированным элементам выборки (посредством наделения их большим весом, что отражается на издержках их неверной классификации). Каждая итерация в конце будет представлять отдельный классификатор. Для финальной классификации прогнозы этих отдельных моделей будут взвешены по их точности. Большинством голосов будет производиться классификация элементов.

Бэггинг был предложен Л. Брайнманом в 1996 г. [23]. Идея данного ансамблирования заключена в создании множества новых выборок данных на основе имеющейся случайным отбором элементов с возвращением. Каждая новая выборка при условии достаточного размера исходной будет содержать лишь 63% уникальных значений начальной выборки [24]. Иными словами, созданные выборки будут сильно отличаться от исходной, а обученные на них классификаторы будут в значительной степени диверсифицированы между собой. Соединение их прогнозов (взвешенным голосованием или по большинству) будет давать более устойчивые и обобщенные (не переобученные) результаты.

Стекинг как ансамблевое решение больше фокусируется на этапе объединения прогнозов отдельных классификаторов. Обычно обучаются наиболее отличные друг от друга как по использованным данным, так и по природе классификаторы (так называемые, базовые модели) (логистическая регрессия, деревья решений, бустинг, лес случайных деревьев). В конце их прогнозы выступают переменными для моделирования финальной модели, так называемого стэка. Как правило, стэком становится логистическая регрессия. Д. Волперт показал, что стэкинг обычно

демонстрирует большую точность классификации нежели любой из базовых моделей (которые входят в него как переменные) [25].

В наши дни ансамбли стали доминирующей схемой в анализе данных, если судить по доле их использования участниками и победителями соревновательных платформ вроде kaggle.com. Как отмечает Н. Оза, широту спектра областей приложения ансамблевых решений можно представить хотя бы по тому, как часто они используются в исторически наиболее трудных задачах классификации: дистанционное зондирование, распознавание личности, распознавание «один против всех» и медицина [26].

Тем не менее ансамблевые решения изначально не предназначены для преодоления проблемы несбалансированности классов, их главная цель — повышение точности классификации. Однако были предложены многочисленные дополнения и модификации ансамблирования, которые позволяют им успешно преодолевать проблему несбалансированности классов в выборочных данных (*рис. 5*).

В первую очередь возможно предварительно обработать данные до обучения (похоже на группу «уровень данных» из *рис. 1*). Подходы здесь группируются по лежащему в основе алгоритму ансамблирования.

Ансамбли на основе бустинга изменяют распределение весов элементов выборки таким образом, чтобы на каждой последующей итерации большее внимание уделялось качеству классификации миноритарного класса. SMOTEBoost на каждой итерации обучения добавляет синтетические элементы меньшего класса посредством метода SMOTE [27]. Новые элементы получают пропорциональные размеру новой выборки веса, а прежние, которые изменяются от итерации к итерации, нормализируются, чтобы составлять вместе с новыми единичную сумму. Аналогичная процедура реализуется в вариации MSMOTEBoost за исключением использования метода MSMOTE для создания новых элементов [28]. RUSBoost случайным образом удаляет представителей доминирующего класса на каждой итерации [29]. Поэтому необходимо лишь нормализировать сумму весов оставшихся элементов выборки.

Подходы на основе бэггинга отличаются большей простотой реализации, поскольку суть алгоритма — в создании новых выборок на основе исходной. Таким образом, соединение бэггинга с методами обработки данных не предполагает каких-либо изменений формул расчета весов элементов выборки или этапов алгоритма. Фокус данных гибридных подходов направлен на то, как создать новые копии выборки, чтобы, с одной стороны, была решена проблема несбалансированности, а с другой стороны, было выполнено условие разнородности выборок. В зависимости от характера предварительной обработки выборочных данных возможно выделить несколько групп таких подходов.

В первую очередь стоит отметить подходы по расширению меньшего класса. Ванг предложил SMOTEBagging и сравнил его с классическим подходом случайного увеличения доли меньшего класса [30]. Классический подход реализуется посредством формирования новой выборки с учетом принадлежности к классам случайно отбираемых элементов. Так, можно полностью сохранить представителей доминирующего класса для каждой новой выборки, а меньший класс отбирать случайным образом до необходимого количества. Либо возможно отбирать и представителей доминирующего класса, чтобы обеспечить большую разнородность выборок. SMOTEBagging отличается не только подходом к расширению доли класса, но и тем, что каждая новая выборка формируется отлично от других. Для каждой выборки выбирается доля от 10 до 100%, часть представителей меньшего класса случайно отбирается из исходной выборки, а оставшееся количество синтезируется посредством SMOTE.

Вторая группа подходов нацелена на уменьшение доли большего класса.

Классический подход предполагает после формирования новой выборки бэггингом случайным образом удаление части представителей. Разные вариации этого подхода были предложены Д. Тао и др. (Asymmetric Bagging) [31] и Е. Чангом и др. (QuasiBagging) [32]. С. Хидо и др. разработали новый метод, в котором количество представителей меньшего класса фиксируется, а количество представителей противоположного класса изменяется от итерации к итерации, следуя отрицательному биномиальному распределению [33]. Более специфический подход предполагает разбиение множества элементов на непересекающиеся подмножества и обучение классификаторов на них [34, 35].

Группа гибридных подходов предполагает одновременное использование и бэггинга, и бустинга, а также подходов предварительной обработки данных. К.-Й. Лиу предложил два варианта — EasyEnsemble и BalanceCascade [36]. Оба подхода используют бэггинг в качестве основного метода, а при обучении каждого отдельного классификатора — бустинг (AdaBoost). Таким образом, финальный алгоритм представляет собой ансамбль ансамблей. Предварительная обработка выборки заключается в снижении доли доминирующего класса. BalanceCascade отличается от EasyEnsemble обработкой представителей большего класса: верно классифицированные представители с высоким уровнем уверенности (confidence level) удаляются из выборки для всех дальнейших итераций бэггинга.

Последнее направление работ по устранению несбалансированности классов посредством ансамблирования (*рис. 5*) объединяет подходы, которые добавляют в стандартные алгоритмы ансамблирования учет издержек неправильной классификации (разный для меньшего и большего классов). Данные алгоритмы основываются на алгоритме AdaBoost и предлагают различные варианты учета издержек в формулу расчета весов элементов. Таковыми являются AdaCost, предложенный В. Фэном и др. [37], CSB1 и CSB2, предложенные К. Тингом и др. [38], а также AdaC1-3, предложенные Й. Саном и др. [39]. М. Джоши и др. разработали RareBoost, который изменяет расчет весов на каждой итерации AdaBoost на основе матрицы ошибок (confusion matrix), содержащей статистику верно и ошибочно классифицированных элементов [40].

В научной литературе можно встретить фундаментальные исследования по сопоставлению тех или иных подходов, основные из которых были упомянуты ранее.

Ряд работ посвящен анализу способности подходов «на уровне данных» по устранению проблемы несбалансированности. Н. Чавла и др. рассмотрели эффективность подходов случайного уменьшения, либо увеличения количества элементов, а также метод SMOTE перед обучением посредством C4.5, наивного байесовского классификатора, и правил принятия решений (Ripper) [7]. Авторы отмечают превосходство SMOTE в преодолении данной проблемы. Э. Истабрукс и др. сравнивают подходы по уменьшению либо увеличению количества элементов и отмечают превосходство подходов, комбинирующих данные методы [41]. Дж. Стефановски и др. сравнивают авторский метод с подходами NCR и SMOTE, делая вывод о наибольшей эффективности предложенного метода и SMOTE [42]. Г. Батиста и др. предлагают исследование спектра методов (10 подходов, три из которых предложены автором) на 13 наборах данных [43]. Авторы отмечают превосходство по критерию точности полученного классификатора предложенных подходов: SMOTE+Tomek, SMOTE+ENN, а также отмечают высокую эффективность достаточно просто метода (случайное увеличение доли меньшего класса). А. Фернандес и др. показывают в своей работе необходимость использования предварительной обработки данных в случае использования метода нечетких правил для классификации [44].

В работах, авторы которых предлагали различные подходы на основе ансамблирования, также сопоставлялись вновь предложенные методы с существующими. Тем не менее эффективность

предлагаемых ими методов оценивалась на данных из конкретных областей деятельности. Очевидно, требуется более объективное сопоставление таких подходов, например, на нескольких наборах данных. Одной из таких работ выступает исследование М. Галара и др. [45]. Авторы сопоставили 22 подхода на основе ансамблирования на 44 наборах данных. Наилучшие результаты показали самые простые подходы: случайное уменьшение доминирующего класса вместе с бэггингом и бустингом. Также отмечается, что большинство алгоритмов на основе ансамблирования (модифицированных для преодоления несбалансированности) превосходят подходы простой предварительной обработки данных («уровень данных» *рис. 1*). Стоит отметить, что из последних был рассмотрен только C4.5, сочетаемый с SMOTE.

Можно заметить, что большинство сравнительных исследований выполнено в отношении небольшого числа методов, а также на специфических данных из разных областей. Тем не менее отсутствуют исследования эффективности подходов простой предварительной обработки данных («уровень данных» *рис. 1*), которые обычно являются первым выбором для банковских данных о кредитовании физических лиц.

В данной работе сделана попытка сопоставления всех изложенных в литературе подходов «на уровне данных» по устранению несбалансированности классов в выборочных данных. Также сопоставлены композиции этих подходов «на уровне данных» в сочетании с ансамблевым решением (стэкинг), что ранее не рассматривалось в научных работах. Сопоставления осуществлены на банковских данных о кредитовании физических лиц.

Собственное исследование

Для проведения исследования будут взяты данные об 278 тыс. кредитов наличными физическим лицам. Кредитование осуществлялось в одной из стран СНГ банком топ-15 по размеру активов по состоянию на 1 октября 2018 г. Изначальный уровень дефолтности данного портфеля составляет 8%, что не позволяет говорить о наличии проблемы несбалансированности данных в сравнении со средними значениями в российской банковской сфере. В силу этого уровень дефолтности искусственно был снижен до 2% посредством произвольного удаления случаев социального дефолта заемщика.

Данные содержат информацию из трех источников: анкетные данные, данные Бюро кредитных историй (БКИ), а также данные о предыдущей внутренней кредитной истории в банке (КИ). Анкетные данные содержат статичную информацию, которую все клиенты представляют в процессе подачи заявки на ссуду. Данные БКИ представляют информацию обо всех имеющихся в Бюро фактах выдачи кредитов другими финансовыми организациями. Кроме того, данные содержат помесячные балансы обслуживания взятых ссуд. Данные КИ представляют информацию о предыдущих заявках на ссуды от клиентов в данном банке. Кроме того, имеются данные по взятым кредитам наличными, потребительским кредитам и кредитным картам. По этим ссудам имеются месячные балансы. Также представлена информация по обслуживанию выставленных платежей.

В рамках данной работы будут сопоставлены все основные подходы на уровне данных (*рис. 2* и *3*) для устранения проблемы несбалансированности классов в выборочных данных. Изменение соотношения классов будет затрагивать только часть исходной выборки (train sample), на которой происходит построение модели (бинарного классификатора). Схема обучения алгоритма классификации представлена на *рис. 6*. Данная схема справедлива для всех рассмотренных подходов, которые опираются на те или иные методы кластеризации, так как присутствует необходимый для кластеризации этап стандартизации данных. Для остальных подходов справедлива схема *рис. 7*.

Последний этап (*рис. 6*) отражает наш подход к оценке качества полученной модели. В качестве примера рассмотрим оценку качества логистической регрессии на данных БКИ для различных уровней дефолта в выборочных данных (искусственное снижение случайным изъятием с 8% до 2% и 4%) —

табл. 1. Разбивка кросс-валидации касается выборок построения (train) и тестирования (valid): все наблюдения, не относящиеся к отложенной выборке (out-of-time, oot), на каждой разбивке случайно распределяются между выборками построения и тестирования. Таким образом, отобранные предикторы дефолта кредитного требования получают перевзвешивание коэффициентов в модели для каждой разбивки кросс-валидации. Соответственно, пересчитываются и оценки качества модели на трех выборках, что дает представление о действительном уровне качества модели и гарантирует не случайность полученных результатов. Стоит отметить, что в *табл. 1* оценки по oot всегда получены на одних и тех же наблюдениях. Итоговые значения для выборки построения — усредненное значение таковых на разбивках кросс-валидации. Для выборки тестирования — полноценная оценка показателя Gini на всех наблюдениях, не относящихся к отложенной выборке, посредством сбора сегментов с разбивок кросс-валидации в единую выборку.

Можно заметить, что проблема несбалансированности классов в выборочных данных действительно влечет негативные эффекты для моделирования сразу по двум характеристикам качества. Во-первых, снижается разделяющая сила модели-классификатора на тестовой выборке (косвенный показатель качества). Так, показатель Gini снижается с 31,3% для оригинального уровня дефолтов (8%) до 30,3% в случае 4-процентного варианта и до 28,7% в случае 2-процентного варианта. Во-вторых, фиксируется снижение качества модели по прямому показателю — обобщающей способности модели, что можно измерить величиной падения показателя Gini между train и oot выборками. Так, величина падения показателя Gini между данными выборками растет с –1,1 п.п. для оригинального уровня дефолтов (8%) до 0,2 п.п. в случае 4-процентного варианта и до 0,8 п.п. в случае 2-процентного варианта.

Все заявленные в *рис. 6* отсечения факторов-предикторов дефолта ссуды имеют одинаковые пороги. Кроме того, процесс построения модели также содержит эвристики, широко используемые на практике, и является автоматическим (не требует вмешательства человека). Все вместе должно обеспечивать объективность сопоставления результатов деятельности различных подходов по устранению несбалансированности.

Как отмечено ранее, при подведении итогов эффективности подходов мы будем обращать внимание не на высокие показатели точности модели на какой-либо выборке (train/test/oot), а на показатель отсутствия пере- или недообучения классификатора — величина падения показателя Gini модели от выборки к выборке (train-test, train-oot). Незначительное (в пределах двух пунктов показателя Gini) падение будет свидетельствовать о достаточной обобщающей способности модели. Иными словами, модель будет с близким к заявленному (на построении) уровню точности классифицировать объекты в будущем.

В качестве классификатора выбрана логистическая регрессия, которая по-прежнему остается основным алгоритмом в моделировании дефолта кредитных требований. Для каждого источника данных будет построена отдельная логистическая регрессия. Для получения финальной модели прогнозные вероятности этих моделей послужат предикторами в рамках стэкинга (ансамбль), который также будет на основе алгоритма логистической регрессии. Таким образом, будет возможность сопоставить не только подходы «на уровне данных», но и группу ансамблевых решений, сочетаемых с предобработкой данных (*рис. 5*). Схема стэкинга представлена на *рис. 8*.

Полученные результаты

В ходе проведения расчетов мы ожидаемо столкнулись с невозможностью задавать целевые уровни дефолтов для ряда подходов (*рис. 9* — выделены серой заливкой), что и предполагалось их авторами. В силу этого мы перебираем различные уровни дефолтности для «контролируемых» подходов (4, 8, 12, 16%) и лучшие варианты для них в дальнейшем будут сопоставлены с результатами «неконтролируемых» подходов.

Результаты падения Gini отсутствуют для подходов и вариантов увеличения

дефолтности выборки построения, если в ходе моделирования не удавалось построить хотя бы одну из моделей по трем источникам данных — по причине отсутствия допущенных переменных-предикторов наложенными ограничениями.

В *табл. 2* можно видеть, как был осуществлен выбор оптимальной дефолтности для «контролируемых» подходов. Главным критерием выступило среднее (по трем моделям источников) падение значения показателя Gini в процентных пунктах между выборкой построения и отложенной выборкой. При прочих равных вторым критерием выступало среднее падение значения показателя Gini в процентных пунктах между выборкой построения и тестовой выборкой. Оптимальные уровни дефолтов отмечены серой заливкой (*табл. 2*). Будем выбирать значения, наиболее близкие по модулю к нулю, так мы избежим и переобучения, и недообучения модели.

Далее возможно сравнить все подходы (*табл. 3*). Как отмечалось ранее, мы допускаем, что при падении показателя Gini до 2 п.п. (по модулю) как между выборкой построения и отложенной выборкой, так и между выборкой построения и тестовой выборкой, можно говорить о достаточной обобщающей способности модели. Среди подходящих под данное ограничение подходов мы будем выбирать по косвенному критерию качества модели — величине показателя Gini на тестовой выборке. Таким образом, наилучшим подходом (из числа подходов «на уровне данных») по устранению несбалансированности классов на данных кредитования наличными оказался EditedNearestNeighbours, а именно: вариант с необходимостью наличия в найденном вокруг наблюдения кластере представителей только того же класса, чтобы наблюдение было оставлено в данных.

Аналогично были сопоставлены ансамблевое решение (стэкинг), сочетаемое с предобработкой данных рассмотренными методами (*табл. 4*). Можно видеть, что наиболее эффективным подходом здесь оказался наиболее простой алгоритм — RandomOverSampler. Данный подход увеличивает численность представителей меньшего класса посредством случайного отбора с возвращением случаев дефолта кредитного требования.

Заключение

В рамках работы была произведена попытка исследования оптимальных подходов к устранению широко представленной в банковской сфере проблеме несбалансированности классов в выборочных данных при моделировании дефолта кредитного требования.

Проблема несбалансированности классов в выборочных данных, как было показано в *табл. 1* нелинейно приводит к снижению потенциала для моделирования: чем меньше дефолтность моделируемого кредитного портфеля, тем хуже будет качество моделей классификации.

Были сопоставлены 21 подход «на уровне данных» по устранению несбалансированности классов в выборочных данных. Также сопоставлены композиции этих подходов «на уровне данных» в сочетании с ансамблевым решением (стэкинг), что ранее не рассматривалось в научных работах. Сопоставления осуществлены на данных кредитования наличными в крупном розничном банке СНГ.

Согласно полученным результатам наилучшие методы среди подходов «на уровне данных» позволили не только восстановить, но и превзойти исходный потенциал для моделирования. Это можно оценить по усредненному (среди отдельных моделей на источниках данных) показателю Gini на исходных данных и в случае использования подхода EditedNearestNeighbours (*табл. 1* и *табл. 3*) — 31,3% против 33,7%.

Стоит отметить, что в случае самостоятельного использования подходов (подходы «на уровне данных»), наилучшие результаты достигаются довольно сложным алгоритмом (EditedNearestNeighbours). В то же время при сочетании ансамблевого решения с данными подходами наилучшие результаты достигаются наиболее простым алгоритмом (RandomOverSampler).

Таблица 1
Снижение качества моделей в силу несбалансированности классов, %

Table 1
The deterioration of the quality of models due to the class imbalance, percent

Разбивка кросс-валидации	2% уровень дефолтов			4% уровень дефолтов			8% уровень дефолтов		
	train	valid	oot	train	valid	oot	train	valid	oot
0	29,3	28,1	28,5	30,2	31,9	30,4	31,4	31,8	32,6
1	29,6	27,3	28,5	31,2	28,4	30,5	31,2	32,3	32,6
2	28,7	30,7	28,2	30,8	29,5	30,5	31,9	29,6	32,6
3	28,6	30,9	28,3	30,1	32,6	30,4	31,4	31,5	32,5
4	29,7	26,7	28,5	30,9	29,2	30,4	31,4	31,5	32,6
Итого	***29,2%***	***28,7***	***28,4***	***30,6***	***30,3***	***30,4***	***31,5***	***31,3***	***32,6***

Источник: авторская разработка

Source: Authoring

Таблица 2
Выбор наилучшего уровня дефолтности для «контролируемых» подходов

Table 2
Selecting the most appropriate default level of *controllable* approaches

Подходы	4%		8%		12%		16%	
	train-test	train-oot	train-test	train-oot	train-test	train-oot	train-test	train-oot
Исходные данные	—	—	0,12	0,01	—	—	—	—
RandomOverSampler	0,65	0,19	0,77	–0,45	0,68	–0,29	0,64	–0,34
SMOTE	3,75	2,51	6,14	4,9	4,29	2,84	3,59	3,34
ADASYN	4,21	2,74	6,34	4,76	4,24	1,93	3,97	3,24
BorderlineSMOTE-1	6,94	4,69	3,7	2,4	—	—	—	—
BorderlineSMOTE-2	4,14	3,27	—	—	—	—	—	—
KmeansSMOTE	7,02	7,91	—	—	—	—	—	—
SVMSMOTE	—	—	—	—	—	—	—	—
RandomUnderSampler	0,57	–0,71	0,56	–0,27	0,52	–0,24	0,71	0,15
ClusterCentroids Soft	—	—	—	—	—	—	—	—
ClusterCentroids Hard	—	—	—	—	—	—	—	—
NearMiss-1	7,32	6,16	—	—	—	—	—	—
NearMiss-2	6,9	6,93	—	—	—	—	—	—
NearMiss-3	–0,79	–0,68	—	—	—	—	—	—

Источник: авторская разработка

Source: Authoring

Таблица 3
Сопоставление всех рассмотренных подходов

Table 3
The comparison of all analyzable approaches

Подходы	**train-test**	**train-oot**	**Gini**		
			train	**test**	**oot**
Исходные данные	0,12	0,01	34,08	33,96	34,08
RandomOverSampler	0,65	0,19	33,51	32,86	33,32
SMOTE	3,75	2,51	34,23	30,48	31,72
ADASYN	4,24	1,93	30,03	25,79	28,1
BorderlineSMOTE-1	3,7	2,4	9,1	5,4	6,7
BorderlineSMOTE-2	4,14	3,27	7,13	2,98	3,86
KmeansSMOTE	7,02	7,91	25,27	18,26	17,36
SVMSMOTE	—	—	—	—	—
RandomUnderSampler	0,71	0,15	33,67	32,96	33,52
NearMiss-1	7,32	6,16	18,47	11,15	12,3
NearMiss-2	6,9	6,93	23,15	16,25	16,23
NearMiss-3	–0,79	–0,68	10,27	11,07	10,95
ClusterCentroids Soft	—	—	—	—	—
ClusterCentroids Hard	—	—	—	—	—
TomekLinks dict	1,15	0,6	28,35	27,21	27,75
TomekLinks majority	0,62	–0,62	34,02	33,39	34,63
EditedNearestNeighbours all	***1,73***	***0,75***	***35,43***	***33,7***	***34,68***
EditedNearestNeighbours mode	0,54	–0,56	33,14	32,6	33,71
RepeatedEditedNearestNeighbours	1,87	1,33	34,42	32,55	33,09
AllKNN all	2,95	2,73	35,42	32,47	32,69
AllKNN mode	2,02	1,51	34,08	32,07	32,58
InstanceHardnessThreshold	6,67	7,85	16,4	9,72	8,54

Источник: авторская разработка

Source: Authoring

Таблица 4
Сопоставление всех рассмотренных подходов со стэкингом

Table 4
The comparison of all analyzable approaches with stacking

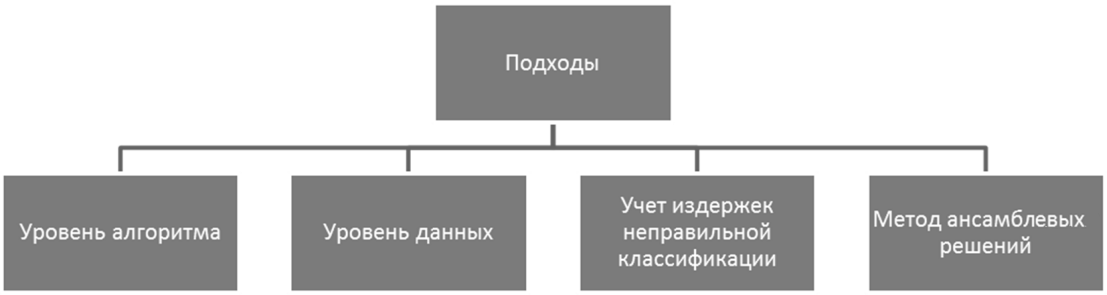
Подходы	train-test	train-oot	Gini		
			train	test	oot
Исходные данные	0,06	0,84	49,18	49,12	48,33
RandomOverSampler	***0,31***	***0,09***	***49,45***	***49,14***	***49,36***
SMOTE	–0,44	0,1	37,92	38,36	37,81
ADASYN	0,09	–2,46	44,62	44,54	47,08
BorderlineSMOTE-1	7,31	2,72	41,97	34,66	39,25
BorderlineSMOTE-2	22,81	22,28	25,17	2,35	2,89
KmeansSMOTE	23,97	24,71	67,39	43,42	42,69
SVMSMOTE	—	—	—	—	—
RandomUnderSampler	0,24	–0,04	49,08	48,84	49,12
NearMiss-1	–2,66	–6,37	30,81	33,47	37,18
NearMiss-2	1,35	1,58	7,75	6,4	6,17
NearMiss-3	10,09	5,68	2,42	–7,67	–3,26
ClusterCentroids Soft	—	—	—	—	—
ClusterCentroids Hard	—	—	—	—	—
TomekLinks dict	–0,64	–1,63	44,53	45,17	46,16
TomekLinks majority	0,99	–1,59	46,42	45,43	48,01
EditedNearestNeighbours all	2,98	–0,05	47,42	44,43	47,46
EditedNearestNeighbours mode	0,64	–0,3	44,57	43,93	44,87
RepeatedEditedNearestNeighbours	3,69	2,59	47,96	44,27	45,37
AllKNN all	3,53	2,3	47,77	44,24	45,46
AllKNN mode	1,41	0,3	45,73	44,33	45,44
InstanceHardnessThreshold	32,33	33,09	45,4	13,06	12,31

Источник: авторская разработка

Source: Authoring

Рисунок 1
Классификация подходов к решению несбалансированности классов

Figure 1
The classification of approaches to addressing the class imbalance

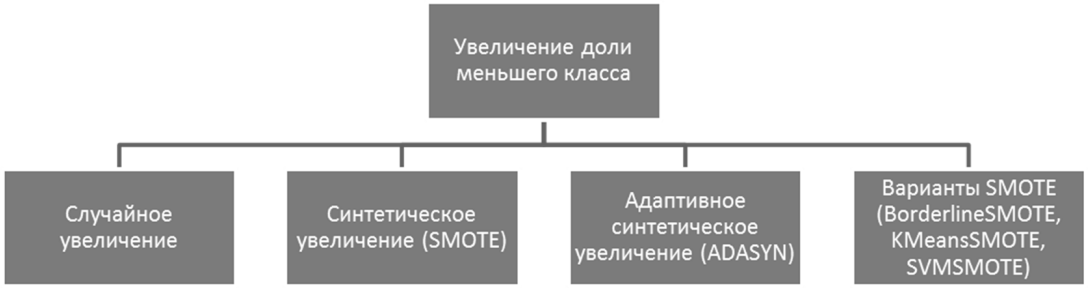


Источник: авторская разработка

Source: Authoring

Рисунок 2
Классификация подходов к увеличению доли меньшего класса

Figure 2
The classification of approaches to increasing the percentage of the minority class

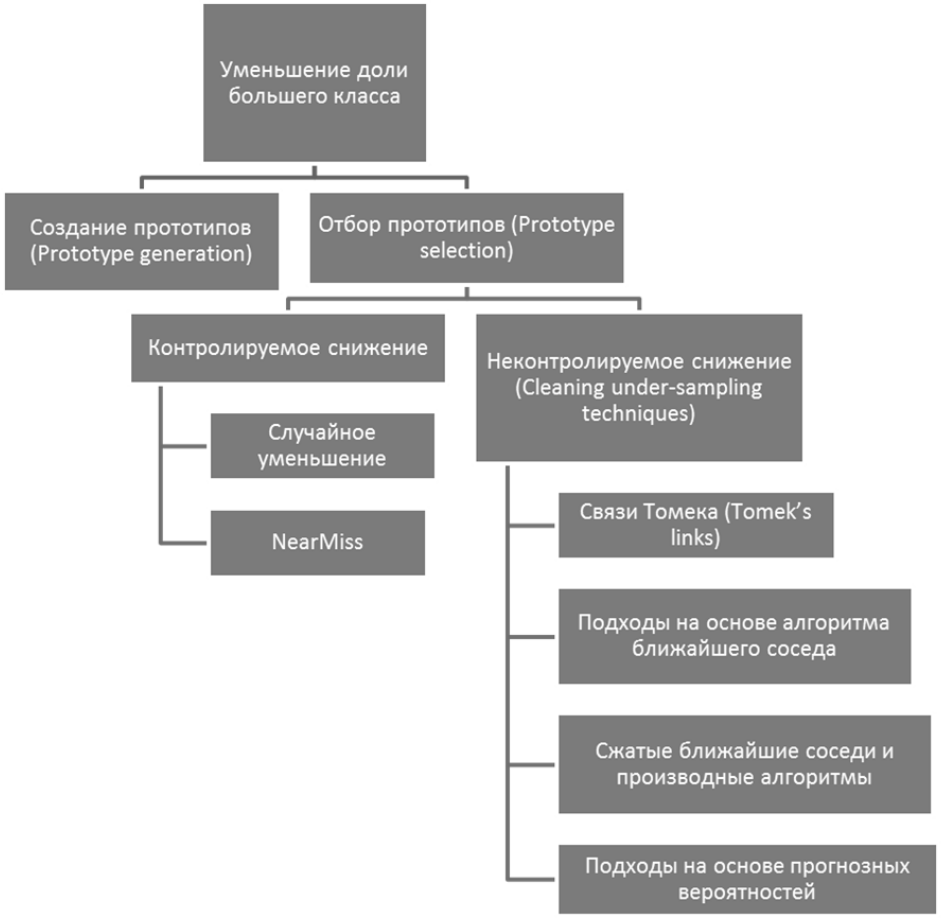


Источник: авторская разработка

Source: Authoring

Рисунок 3
Классификация подходов к уменьшению доли доминирующего класса

Figure 3
The classification of approaches to decreasing the percentage of the dominant class



Источник: авторская разработка

Source: Authoring

Рисунок 4
Классификация видов ансамблирования

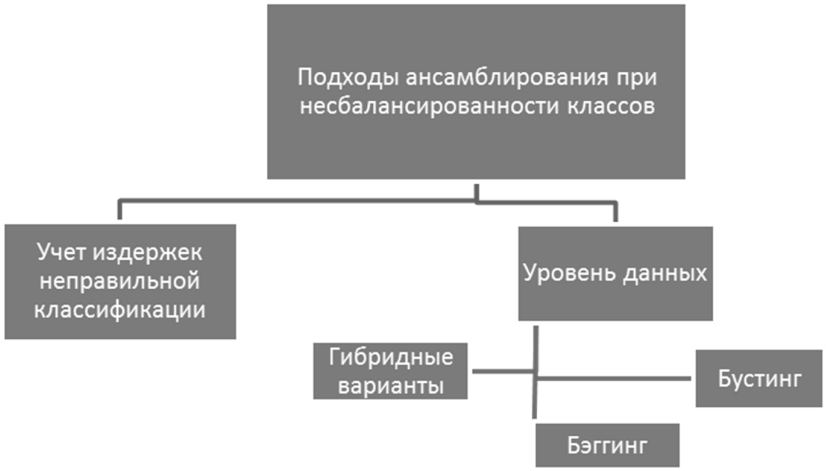
Figure 4
The classification of ensemble types

Источник: авторская разработка

Source: Authoring

Рисунок 5
Классификация подходов на основе ансамблирования

Figure 5
The classification of ensemble-based approaches



Источник: авторская разработка

Source: Authoring

Рисунок 6
Первая схема обучения классификатора

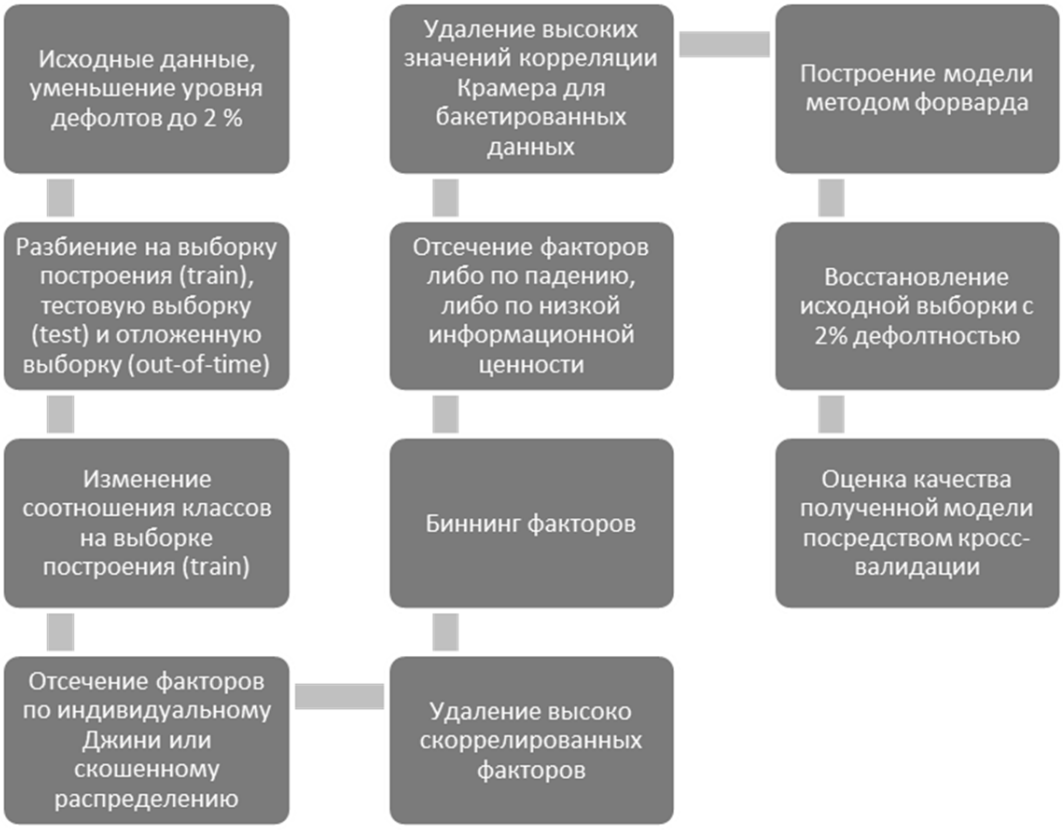
Figure 6
The first classifier training scheme

Источник: авторская разработка

Source: Authoring

Рисунок 7
Вторая схема обучения классификатора

Figure 7
The second classifier training scheme

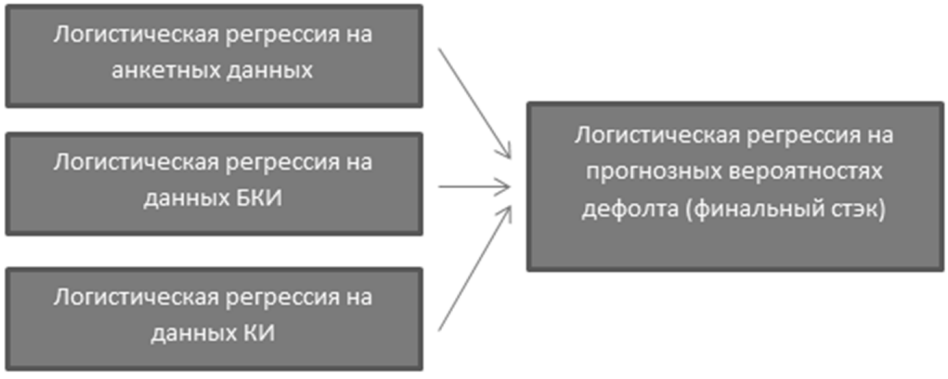


Источник: авторская разработка

Source: Authoring

Рисунок 8
Схема финальной модели прогнозирования дефолта заемщика

Figure 8
The scheme of the final model for the debtor's default forecast

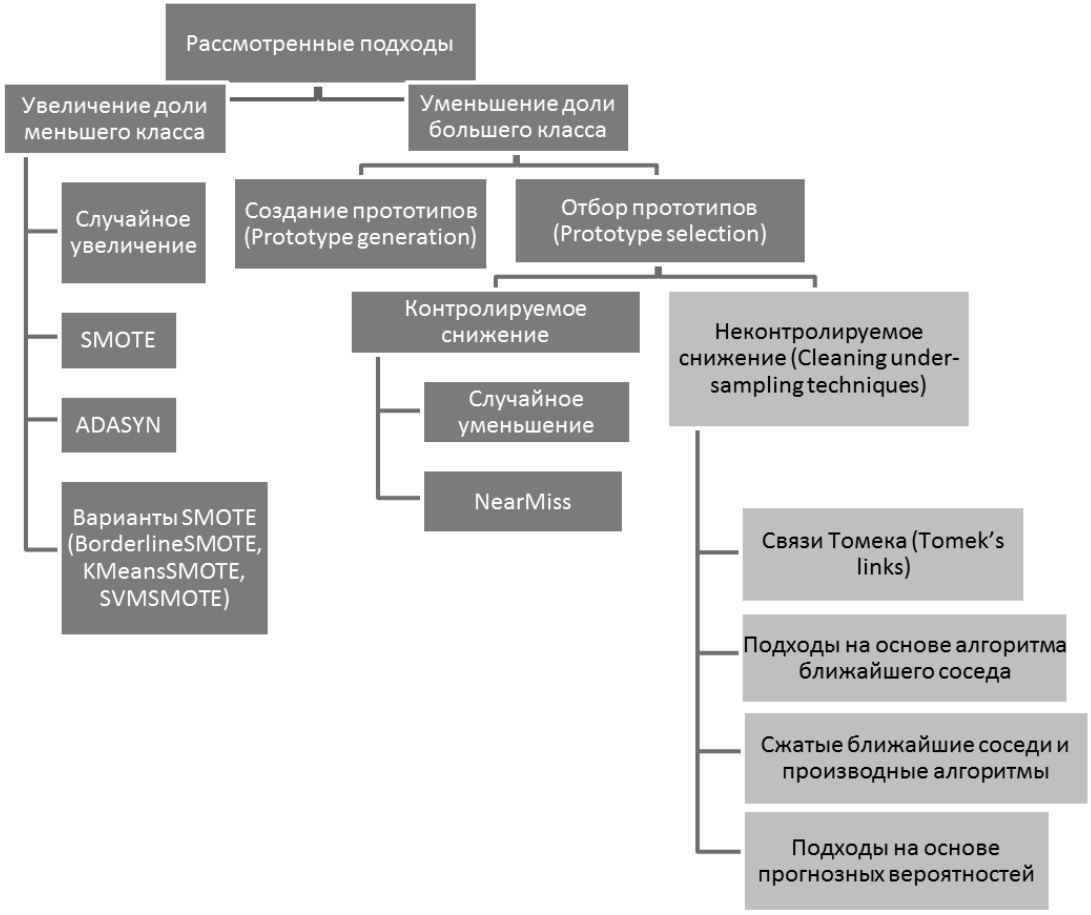


Источник: авторская разработка

Source: Authoring

Рисунок 9
Классификация рассмотренных подходов

Figure 9
The classification of analyzable approaches



Источник: авторская разработка

Source: Authoring

Список литературы

1. *Sun Y., Wong A.C., Kamel M.S.* Classification of imbalanced data: A review. *International Journal of Pattern Recognition and Artificial Intelligenc*e, 2009, vol. 23, no. 4, pp. 687–719. URL: https://doi.org/10.1142/S0218001409007326

2. *García V., Mollineda R., Sánchez J.* On the *k*-NN performance in a challenging scenario of imbalance and overlapping. *Pattern Analysis and Applications*, 2008, vol. 11, iss. 3-4, pp. 269–280. URL: https://doi.org/10.1007/s10044-007-0087-5

3. *Japkowicz N., Stephen S.* The class imbalance problem: A systematic study. *Intelligent Data Analysis*, 2002, vol. 6, no. 5, pp. 429–449 URL: http://dx.doi.org/10.3233/IDA-2002-6504

4. *Weiss G.M., Provost F.* Learning when training data are costly: The effect of class distribution on tree induction. *Journal of Artificial Intelligence Research*, 2003, vol. 19, pp. 315–354. URL: https://doi.org/10.1613/jair.1199

5. *Lin Y., Lee Y., Wahba G.* Support vector machines for classification in nonstandard situations. *Machine Learning*, 2002, vol. 46, iss. 1-3, pp. 191–202. URL: https://doi.org/10.1023/A:1012406528296

6. *Wu G., Chang E.* KBA: kernel boundary alignment considering imbalanced data distribution. *IEEE Transactions on Knowledge and Data Engineering*, 2005, vol. 17, iss. 6, pp. 786–795. URL: https://ieeexplore.ieee.org/document/1423979

7. *Chawla N.V., Bowyer K.W., Hall L.O., Kegelmeyer W.P.* SMOTE: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 2002, vol. 16, pp. 321–357. URL: https://doi.org/10.1613/jair.953

8. *He H., Bai Y., Garcia E.A., Li S.* ADASYN: Adaptive synthetic sampling approach for imbalanced learning. *IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, 2008, pp. 1322–1328. URL: https://ieeexplore.ieee.org/document/4633969/

9. *Han H., Wang W.-Y., Mao B.-H.* Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning. In: Huang D.S., Zhang X.P., Huang G.B. (eds) Advances in Intelligent Computing. ICIC 2005. Lecture Notes in Computer Science, 2005, vol. 3644, pp. 878–887. URL: https://doi.org/10.1007/11538059_91

10. *Nguyen H.M., Cooper E.W., Kamei K.* Borderline over-sampling for imbalanced data classification. *International Journal of Knowledge Engineering and Soft Data Paradigms*, 2011, vol. 3, iss. 1, pp. 4–21. URL: http://www.inderscience.com/offer.php?id=39875

11. *Last F., Douzas G., Bacao F.* Oversampling for Imbalanced Learning Based on *k*-Means and SMOTE. URL: https://arxiv.org/abs/1711.00837

12. *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. URL: https://www.site.uottawa.ca/~nat/Workshop2003/jzhang.pdf

13. *Tomek I.* Two modifications of CNN. *IEEE Transactions on Systems, Man, and Cybernetics*, 1976, vol. SMC-6, iss. 11, pp. 769–772. URL: https://doi.org/10.1109/TSMC.1976.4309452

14. *Wilson D.* Asymptotic Properties of Nearest Neighbor Rules Using Edited Data. *IEEE Transactions on Systems, Man, and Cybernetics*, 1972, vol. SMC-2, iss. 3, pp. 408–421. URL: https://ieeexplore.ieee.org/document/4309137/

15. *Hart P.* The condensed nearest neighbor rule. *IEEE Transactions on Information Theory*, 1968, vol. 14, iss. 3, pp. 515–516. URL: https://ieeexplore.ieee.org/document/1054155/

16. *Kubat M., Matwin S.* Addressing the curse of imbalanced training sets: one-sided selection. *Proceedings of the Fourteenth International Conference on Machine Learning*, 1997, vol. 97, pp. 179–186.

17. *Smith M.R., Martinez T., Giraud-Carrier C.* An instance level analysis of data complexity. *Machine Learning*, 2014, vol. 95, iss. 2, pp. 225–256. URL: https://doi.org/10.1007/s10994-013-5422-z

18. *Domingos P.* MetaCost: a general method for making classifiers cost-sensitive. *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1999, pp. 155–164. URL: https://doi.org/10.1145/312129.312220

19. *Ling C.X., Sheng V.S., Yang Q.* Test strategies for cost-sensitive decision trees. *IEEE Transactions on Knowledge and Data Engineering*, 2006, vol. 18, iss. 8, pp. 1055–1067. URL: https://ieeexplore.ieee.org/document/1644729/

20. *Schapire R.E.* The strength of weak learnability. *Machine Learning*, 1990, vol. 5, iss. 2, pp. 197–227. URL: https://doi.org/10.1007/BF00116037

21. *Freund Y., Schapire R.E.* A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 1997, vol. 55, iss. 1, pp. 119–139. URL: https://doi.org/10.1006/jcss.1997.1504

22. *Schapire R.E., Singer Y.* Improved boosting algorithms using confidence-rated predictions. *Machine Learning*, 1999, vol. 37, iss. 3, pp. 297–336. URL: https://doi.org/10.1023/A:1007614523901

23. *Breiman L.* Bagging predictors. *Machine Learning*, 1996, vol. 24, iss. 2, pp. 123–140. URL: https://doi.org/10.1023/A:1018054314350

24. *Aslam J.A., Popa R.A., Rivest R.L.* On Estimating the Size and Confidence of a Statistical Audit. *Proceedings of the USENIX Workshop on Accurate Electronic Voting Technology*, 2007.

25. *Wolpert D.H.* Stacked Generalization. *Neural Networks*, 1992, vol. 5, iss. 2, pp. 241-259. URL: https://doi.org/10.1016/S0893-6080(05)80023-1

26. *Oza N.C., Tumer K.* Classifier ensembles: Select real-world applications. *Information Fusion*, 2008, vol. 9, iss. 1, pp. 4–20. URL: https://doi.org/10.1016/j.inffus.2007.07.002

27. *Chawla N.V., Lazarevic A., Hall L.O., Bowyer K.W.* SMOTEBoost: Improving prediction of the minority class in boosting. In: Lavrač N., Gamberger D., Todorovski L., Blockeel H. (eds) Knowledge Discovery in Databases: PKDD 2003. *PKDD 2003. Lecture Notes in Computer Science*, 2003, vol. 2838, Berlin, Springer, pp. 107–119. URL: https://doi.org/10.1007/978-3-540-39804-2_12

28. *Seiffert C., Khoshgoftaar T.M., Van Hulse J., Napolitano A.* Rusboost: A hybrid approach to alleviating class imbalance. *IEEE Transactions on Systems, Man, and Cybernetics – Part A: Systems and Humans*, 2010, vol. 40, iss. 1, pp. 185–197. URL: https://doi.org/10.1109/TSMCA.2009.2029559

29. *Hu S., Liang Y., Ma L., He Y.* MSMOTE: Improving classification performance when training data is imbalanced. *Second International Workshop on Computer Science and Engineering*, 2009, vol. 2, pp. 13–17. URL: https://doi.org/10.1109/WCSE.2009.756

30. *Wang S., Yao X.* Diversity analysis on imbalanced data sets by using ensemble models. *IEEE Symposium on Computational Intelligence and Data Mining*, 2009, pp. 324–331. URL: https://ieeexplore.ieee.org/document/4938667/

31. *Tao D., Tang X., Li X., Wu X.* Asymmetric bagging and random subspace for support vector machines-based relevance feedback in image retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2006, vol. 28, iss. 7, pp. 1088–1099. URL: https://doi.org/10.1109/TPAMI.2006.134

32. *Chang E., Li B., Wu G., Goh K.* Statistical learning for effective visual information retrieval. *Proceedings 2003 International Conference on Image Processing*, 2003, pp. 609–612. URL: https://ieeexplore.ieee.org/document/1247318

33. *Hido S., Kashima H., Takahashi Y.* Roughly balanced bagging for imbalanced data. *Statistical Analysis and Data Mining*, 2009, vol. 2, iss. 5-6, pp. 412–426. URL: https://doi.org/10.1002/sam.10061

34. *Chan P.K., Stolfo S.J.* Toward scalable learning with non-uniform class and cost distributions: A case study in credit card fraud detection. *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining*, 1998, pp. 164–168. URL: https://pdfs.semanticscholar.org/6e19/3366945bf3bd72d5ba906e3982ac4d8ae874.pdf

35. *Yan R., Liu Y., Jin R., Hauptmann A.* On predicting rare classes with SVM ensembles in scene classification. *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03)*, 2003, vol. 3, pp. 21–24. URL: https://doi.org/10.1109/ICASSP.2003.1199097

36. *Liu X.-Y., Wu J., Zhou Z.-H.* Exploratory undersampling for class-imbalance learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 2009, vol. 39, iss. 2, pp. 539–550. URL: https://doi.org/10.1109/TSMCB.2008.2007853

37. *Fan W., S. Stolfo J., Zhang J., Chan P.K.* Adacost: Misclassification cost-sensitive boosting. *Proceedings of the Sixteenth International Conference on Machine Learning*, 1999, pp. 97–105.

38. *Ting K.M.* A comparative study of cost-sensitive boosting algorithms. *Proceedings of the Seventeenth International Conference on Machine Learning*, 2000, pp. 983–990.

39. *Sun Y., Kamel M.S., Wong A.K., Wang Y.* Cost-sensitive boosting for classification of imbalanced data. *Pattern Recognition*, 2007, vol. 40, iss. 12, pp. 3358–3378. URL: https://doi.org/10.1016/j.patcog.2007.04.009

40. *Joshi M.V., Kumar V., Agarwal R.C.* Evaluating boosting algorithms to classify rare classes: Comparison and improvements. *Proceedings 2001 IEEE International Conference on Data Mining*, 2001, pp. 257–264. URL: https://ieeexplore.ieee.org/document/989527/

41. *Estabrooks A., Jo T., Japkowicz N.* A multiple resampling method for learning from imbalanced data sets. *Computational Intelligence*, 2004, vol. 20, iss. 1, pp. 18–36. URL: https://doi.org/10.1111/j.0824-7935.2004.t01-1-00228.x

42. *Stefanowski J., Wilk S.* Selective pre-processing of imbalanced data for improving classification performance. In: Song I.Y., Eder J., Nguyen T.M. (eds) Data Warehousing and Knowledge Discovery. DaWaK 2008. *Lecture Notes in Computer Science*, 2008, vol. 5182, Berlin, Springer, pp. 283–292. URL: https://doi.org/10.1007/978-3-540-85836-2_27

43. *Batista G.E.A.P.A., Prati R.C., Monard M.C.* A study of the behavior of several methods for balancing machine learning training data. *SIGKDD Explorations Newsletter*, 2004, vol. 6, iss. 1, pp. 20–29.

44. *Fernandez A., Garcıa S., del Jesus M.J., Herrera F.* A study of the behaviour of linguistic fuzzy rule based classification systems in the framework of imbalanced data-sets. *Fuzzy Sets and Systems*, 2008, vol. 159, iss. 18, pp. 2378–2398. URL: https://doi.org/10.1016/j.fss.2007.12.023

45. *Galar M., Fernandez A., Barrenechea E. et al*. A Review on Ensembles for the Class Imbalance Problem: Bagging-, Boosting-, and Hybrid-Based Approaches. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 2012, vol. 42, iss. 4, pp. 463–484. URL: https://doi.org/10.1109/TSMCC.2011.2161285

Информация о конфликте интересов

Я, автор данной статьи, со всей ответственностью заявляю о частичном и полном отсутствии фактического или потенциального конфликта интересов с какой бы то ни было третьей стороной, который может возникнуть вследствие публикации данной статьи. Настоящее заявление относится к проведению научной работы, сбору и обработке данных, написанию и подготовке статьи, принятию решения о публикации рукописи.

pISSN 2071-4688
eISSN 2311-8709

Banking

OVERCOMING THE CLASS IMBALANCE IN MODELING THE CREDIT DEFAULT

Vladislav Vl. ROSKOSHENKO

Lomonosov Moscow State University (MSU), Moscow, Russian Federation
roskoshenkoeco@mail.ru
ORCID: not available

Article history:
Article No. 667/2019
Received 17 October 2019
Received in revised form
31 October 2019
Accepted 14 November 2019
Available online
29 November 2019

JEL classification: G21, G28

Keywords: credit scoring, logistic regression, ensemble, class imbalance, binary classification

Abstract

Subject The banking sector faces the class imbalance of samples in modeling the credit default. Data pre-processing is traditionally the first option to choose in bank modeling, since it helps overcome the class imbalance. Available studies into such approaches and their comparison discuss a few methods or focus on very specific data. Moreover, previous researchers overlook approaches combining data pre-processing and ensemble-based solutions (stacking).

Objectives The study aims to find the best-fit option to overcome the class imbalance of each group of approaches applied to bank data on retail lending.

Methods The study employs mathematical modeling, statistical analysis and content analysis of sources.

Results Although being rather mathematically difficult, EditedNearestNeighbours approach proved to be most convenient for pre-processing of data. It excludes representatives of the dominant class, which are inadequate to the surrounding environment which is determined through clustering. RandomOverSampler also turned to meet expectations among combinations of data pre-processing and stacking approaches. It increases a percentage of the minority class randomly and appears to be most simple.

Conclusions and Relevance The article presents an exhaustive comparison of approaches to the class imbalance in samples. I selected the most appropriate approach from data pre-processing approaches and the best combination of data pre-processing and ensemble-based solution. The findings can be used for purposes of credit scoring and statistical modeling, when binary classification is required.

Please cite this article as: Roskoshenko V.Vl. Overcoming the Class Imbalance in Modeling the Credit Default. *Finance and Credit*, 2019, vol. 25, iss. 11, pp. 2534–2561.
https://doi.org/10.24891/fc.25.11.2534

References

1. Sun Y., Wong A.C., Kamel M.S. Classification of Imbalanced Data: A Review. *International Journal of Pattern Recognition and Artificial Intelligenc*e, 2009, vol. 23, no. 4, pp. 687–719. URL: https://doi.org/10.1142/S0218001409007326

2. García V., Mollineda R., Sánchez J. On the *k*-NN Performance in a Challenging Scenario of Imbalance and Overlapping. *Pattern Analysis and Applications*, 2008, vol. 11, iss. 3-4, pp. 269–280. URL: https://doi.org/10.1007/s10044-007-0087-5

3. Japkowicz N., Stephen S. The Class Imbalance Problem: A Systematic Study. *Intelligent Data Analysis*, 2002, vol. 6, no. 5, pp. 429–449. URL: http://dx.doi.org/10.3233/IDA-2002-6504

4. Weiss G.M., Provost F. Learning When Training Data are Costly: The Effect of Class Distribution on Tree Induction. *Journal of Artificial Intelligence Research*, 2003, vol. 19, pp. 315–354. URL: https://doi.org/10.1613/jair.1199

5. Lin Y., Lee Y., Wahba G. Support Vector Machines for Classification in Nonstandard Situations. *Machine Learning*, 2002, vol. 46, iss. 1-3, pp. 191–202. URL: https://doi.org/10.1023/A:1012406528296

6. Wu G., Chang E. KBA: Kernel Boundary Alignment Considering Imbalanced Data Distribution. *IEEE Transactions on Knowledge and Data Engineering*, 2005, vol. 17, iss. 6, pp. 786–795. URL: https://ieeexplore.ieee.org/document/1423979

7. Chawla N.V., Bowyer K.W., Hall L.O., Kegelmeyer W.P. SMOTE: Synthetic Minority Over-Sampling Technique. *Journal of Artificial Intelligence Research*, 2002, vol. 16, pp. 321–357. URL: https://doi.org/10.1613/jair.953

8. He H., Bai Y., Garcia E.A., Li S. ADASYN: Adaptive Synthetic Sampling Approach for Imbalanced Learning. *IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, 2008, pp. 1322–1328. URL: https://ieeexplore.ieee.org/document/4633969/

9. Han H., Wang W.-Y., Mao B.-H. Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning. In: Huang DS., Zhang XP., Huang GB. (eds) Advances in Intelligent Computing. ICIC 2005. Lecture Notes in Computer Science, 2005, vol. 3644, pp. 878–887. URL: https://doi.org/10.1007/11538059_91

10. Nguyen H.M., Cooper E.W., Kamei K. Borderline Over-Sampling for Imbalanced Data Classification. *International Journal of Knowledge Engineering and Soft Data Paradigms*, 2011, vol. 3, iss. 1, pp. 4–21. URL: http://www.inderscience.com/offer.php?id=39875

11. Last F., Douzas G., Bacao F. Oversampling for Imbalanced Learning Based on *k*-Means and SMOTE. URL: https://arxiv.org/abs/1711.00837

12. 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. URL: https://www.site.uottawa.ca/~nat/Workshop2003/jzhang.pdf

13. Tomek I. Two Modifications of CNN. *IEEE Transactions on Systems, Man, and Cybernetics*, 1976, vol. SMC-6, iss. 11, pp. 769–772. URL: https://doi.org/10.1109/TSMC.1976.4309452

14. Wilson D. Asymptotic Properties of Nearest Neighbor Rules Using Edited Data. *IEEE Transactions on Systems, Man, and Cybernetics*, 1972, vol. SMC-2, iss. 3, pp. 408–421. URL: https://ieeexplore.ieee.org/document/4309137/

15. Hart P. The Condensed Nearest Neighbor Rule. *IEEE Transactions on Information Theory*, 1968, vol. 14, iss. 3, pp. 515–516. URL: https://ieeexplore.ieee.org/document/1054155/

16. Kubat M., Matwin S. Addressing the Curse of Imbalanced Training Sets: One-Sided Selection. *Proceedings of the Fourteenth International Conference on Machine Learning*, 1997, vol. 97, pp. 179–186.

17. Smith M.R., Martinez T., Giraud-Carrier C. An Instance Level Analysis of Data Complexity. *Machine Learning*, 2014, vol. 95, iss. 2, pp. 225–256. URL: https://doi.org/10.1007/s10994-013-5422-z

18. Domingos P. MetaCost: A General Method for Making Classifiers Cost-Sensitive. *Proceedings of the fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1999, pp. 155–164. URL: https://doi.org/10.1145/312129.312220

19. Ling C.X., Sheng V.S., Yang Q. Test Strategies for Cost-Sensitive Decision Trees. *IEEE Transactions on Knowledge and Data Engineering*, 2006, vol. 18, iss. 8, pp. 1055–1067. URL: https://ieeexplore.ieee.org/document/1644729/

20. Schapire R.E. The Strength of Weak Learnability. *Machine Learning*, 1990, vol. 5, iss. 2, pp. 197–227. URL: https://doi.org/10.1007/BF00116037

21. Freund Y., Schapire R.E. A Decision-Theoretic Generalization of On-line Learning and an Application to Boosting. *Journal of Computer and System Sciences*, 1997, vol. 55, iss. 1, pp. 119–139. URL: https://doi.org/10.1006/jcss.1997.1504

22. Schapire R.E., Singer Y. Improved Boosting Algorithms Using Confidence-Rated Predictions. *Machine Learning*, 1999, vol. 37, iss. 3, pp. 297–336. URL: https://doi.org/10.1023/A:1007614523901

23. Breiman L. Bagging Predictors. *Machine Learning*, 1996, vol. 24, iss. 2, pp. 123–140. URL: https://doi.org/10.1023/A:1018054314350

24. Aslam J.A., Popa R.A., Rivest R.L. On Estimating the Size and Confidence of a Statistical Audit. *Proceedings of the USENIX Workshop on Accurate Electronic Voting Technology*, 2007.

25. Wolpert D.H. Stacked Generalization. *Neural Networks*, 1992, vol. 5, iss. 2, pp. 241–259. URL: https://doi.org/10.1016/S0893-6080(05)80023-1

26. Oza N.C., Tumer K. Classifier Ensembles: Select Real-World Applications. *Information Fusion*, 2008, vol. 9, iss. 1, pp. 4–20. URL: https://doi.org/10.1016/j.inffus.2007.07.002

27. Chawla N.V., Lazarevic A., Hall L.O., Bowyer K.W. SMOTEBoost: Improving Prediction of the Minority Class in Boosting. In: Lavrač N., Gamberger D., Todorovski L., Blockeel H. (eds) Knowledge Discovery in Databases: PKDD 2003. PKDD 2003. Lecture Notes in Computer Science, 2003, vol. 2838. Berlin, Springer, pp. 107–119. URL: https://doi.org/10.1007/978-3-540-39804-2_12

28. Seiffert C., Khoshgoftaar T.M., Van Hulse J., Napolitano A. Rusboost: A Hybrid Approach to Alleviating Class Imbalance. *IEEE Transactions on Systems, Man, and Cybernetics – Part A: Systems and Humans*, 2010, vol. 40, iss. 1, pp. 185–197. URL: https://doi.org/10.1109/TSMCA.2009.2029559

29. Hu S., Liang Y., Ma L., He Y. MSMOTE: Improving Classification Performance When Training Data is Imbalanced. *Second International Workshop on Computer Science and Engineering*, 2009, vol. 2, pp. 13–17. URL: https://doi.org/10.1109/WCSE.2009.756

30. Wang S., Yao X. Diversity Analysis on Imbalanced Data Sets by Using Ensemble Models. *IEEE Symposium on Computational Intelligence and Data Mining*, 2009, pp. 324–331. URL: https://ieeexplore.ieee.org/document/4938667/

31. Tao D., Tang X., Li X., Wu X. Asymmetric Bagging and Random Subspace for Support Vector Machines-Based Relevance Feedback in Image Retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2006, vol. 28, iss. 7, pp. 1088–1099. URL: https://doi.org/10.1109/TPAMI.2006.134

32. Chang E., Li B., Wu G., Goh K. Statistical Learning for Effective Visual Information Retrieval. *Proceedings 2003 International Conference on Image Processing*, 2003, pp. 609–612. URL: https://ieeexplore.ieee.org/document/1247318

33. Hido S., Kashima H., Takahashi Y. Roughly Balanced Bagging for Imbalanced Data. *Statistical Analysis and Data Mining*, 2009, vol. 2, iss. 5-6, pp. 412–426. URL: https://doi.org/10.1002/sam.10061

34. Chan P.K., Stolfo S.J. Toward Scalable Learning with Non-Uniform Class and Cost Distributions: A Case Study in Credit Card Fraud Detection. *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining*, 1998, pp. 164–168. URL: https://pdfs.semanticscholar.org/6e19/3366945bf3bd72d5ba906e3982ac4d8ae874.pdf

35. Yan R., Liu Y., Jin R., Hauptmann A. On Predicting Rare Classes with SVM Ensembles in Scene Classification. *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings (ICASSP '03)*, 2003, vol. 3, pp. 21–24. URL: https://doi.org/10.1109/ICASSP.2003.1199097

36. Liu X.-Y., Wu J., Zhou Z.-H. Exploratory Undersampling for Class Imbalance Learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 2009, vol. 39, iss. 2, pp. 539–550. URL: https://doi.org/10.1109/TSMCB.2008.2007853

37. Fan W., S. Stolfo J., Zhang J., Chan P.K. Adacost: Misclassification Cost-Sensitive Boosting. *Proceedings of the Sixteenth International Conference on Machine Learning*, 1999, pp. 97–105.

38. Ting K.M. A Comparative Study of Cost-Sensitive Boosting Algorithms. *Proceedings of the Seventeenth International Conference on Machine Learning*, 2000, pp. 983–990.

39. Sun Y., Kamel M.S., Wong A.K., Wang Y. Cost-Sensitive Boosting for Classification of Imbalanced Data. *Pattern Recognition*, 2007, vol. 40, iss. 12, pp. 3358–3378. URL: https://doi.org/10.1016/j.patcog.2007.04.009

40. Joshi M.V., Kumar V., Agarwal R.C. Evaluating Boosting Algorithms to Classify Rare Classes: Comparison and Improvements. *Proceedings 2001 IEEE International Conference on Data Mining*, 2001, pp. 257–264. URL: https://ieeexplore.ieee.org/document/989527/

41. Estabrooks A., Jo T., Japkowicz N. A Multiple Resampling Method for Learning from Imbalanced Data Sets. *Computational Intelligence*, 2004, vol. 20, iss. 1, pp. 18–36. URL: https://doi.org/10.1111/j.0824-7935.2004.t01-1-00228.x

42. Stefanowski J., Wilk S. Selective Pre-Processing of Imbalanced Data for Improving Classification Performance. In: Song IY., Eder J., Nguyen T.M. (eds) Data Warehousing and Knowledge Discovery. DaWaK 2008. Lecture Notes in Computer Science, 2008, vol. 5182. Berlin, Springer, pp. 283–292. URL: https://doi.org/10.1007/978-3-540-85836-2_27

43. Batista G.E.A.P.A., Prati R.C., Monard M.C. A Study of the Behavior of Several Methods for Balancing Machine Learning Training Data. *ACM SIGKDD Explorations Newsletter*, 2004, vol. 6, iss. 1, pp. 20–29.

44. Fernandez A., Garcıa S., del Jesus M.J., Herrera F. A Study of the Behaviour of Linguistic Fuzzy Rule Based Classification Systems in the Framework of Imbalanced Data-sets. *Fuzzy Sets and Systems*, 2008, vol. 159, iss. 18, pp. 2378–2398. URL: https://doi.org/10.1016/j.fss.2007.12.023

45. Galar M., Fernandez A., Barrenechea E. et al. A Review on Ensembles for the Class Imbalance Problem: Bagging-, Boosting-, and Hybrid-Based Approaches. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 2012, vol. 42, iss. 4, pp. 463–484. URL: https://doi.org/10.1109/TSMCC.2011.2161285

Conflict-of-interest notification

I, the author of this article, bindingly and explicitly declare of the partial and total lack of actual or potential conflict of interest with any other third party whatsoever, which may arise as a result of the publication of this article. This statement relates to the study, data collection and interpretation, writing and preparation of the article, and the decision to submit the manuscript for publication.