DOI: 10.15593/2224-9397/2020.1.08
УДК 004.67

Л.А. Мыльников, А.С. Морозов, Д.В. Пухарева

Пермский национальный исследовательский политехнический университет,
Пермь, Россия

ВЫБОР МЕТОДОВ КЛАССИФИКАЦИИ И ПОВЫШЕНИЕ ИХ ЭФФЕКТИВНОСТИ В ЗАДАЧАХ ИДЕНТИФИКАЦИИ НА ПРИМЕРЕ ВЫЯВЛЕНИЯ МОШЕННИКОВ В МАГАЗИНАХ ПОЛНОГО САМООБСЛУЖИВАНИЯ

Актуальность рассматриваемой в статье задачи связана с широким использованием эмпирических моделей и методов классификации для подготовки и предварительной обработки данных и повышения на их основе объективности принимаемых решений за счет учета особенностей рассматриваемых систем и задач на основе характеризующих их статистических данных. **Целью** работы является рассмотрение задачи выбора метода идентификации для повышения эффективности его работы для конкретной прикладной задачи в условиях непрерывного дополнения данных. Для этого в статье рассматривается применение методов машинного обучения на данных статистики, собираемых в режиме реального времени, о действиях пользователей магазинов полного цикла самообслуживания с устройством сканирования товарных штрих-кодов. **Полученные** результаты позволяют классифицировать данные на две категории (идентифицировать интересующие состояния). На рассматриваемом в статье примере это выявление мошенников на основе их действий. Выбранные модели и способы повышения их эффективности могут быть использованы напрямую в тех сферах, где необходим контроль персонала и клиентов на основе их электронных следов в реальном времени. Наибольшая **значимость исследования** связана с тем, что при рассмотрении разных задач методы классификации показывают различную эффективность. В проведенном исследовании **представлена методика** выбора и повышения эффективности моделей для решения задач бинарной классификации и идентификации. Этот процесс сведен к последовательности формальных операций, которые могут быть проведены при решении любой задачи классификации. Эффективность оценивалась с использованием ROC-кривых, а для повышения эффективности работы методов машинного обучения применялись такие подходы, как построение ансамблей моделей, кросс-валидация, использование специальных метрик при обучении моделей и ресемплинг.

Ключевые слова: ритэйл, идентификация, бинарная классификация, машинное обучение, мошенничество, выбор модели, повышение эффективности, объединение моделей, поддержка принятия решений.

L.A. Mylnikov, A.S. Morozov, D.V. Pukhareva

Perm National Research Polytechnic University, Perm, Russian Federation

SELECTING A CLASSIFICATION METHODS AND INCREASING ITS EFFICIENCY IN IDENTIFICATION TASKS ON THE EXAMPLE OF IDENTIFYING FRAUDS IN WALK OUT RETAIL STORES

The **relevance** of the task considered in the paper is connected with the use of empirical models and classification methods for the preprocessing and preparation of data, and on their basis, increasing of accuracy and objectivity of decisions made as well as with the consideration of observed systems and tasks, and the use of statistical data. The paper is **aimed** at considering the identification method selection task to increase its efficiency performance on the base of a specific applied task under continuous data include. For the set task to be handled, the **paper investigates** the application of machine learning methods on the base of real time-collected statistical data about customer actions in self-service retail shops with a special device for scanning bar codes of goods. **The obtained results** allow classifying data into two categories (i.e. to identify target states. In the paper, this is the identification of frauds on the base of customer actions. The selected models and the ways to improve their efficiency can be used directly in the areas where staff and clients' supervision and control in real-time (i.e. based on their e-actions) are of importance. The significance of the study is greatly connected with the outcome according to which classification methods used to consider various tasks show different efficiency levels. In the conducted study is given the method which determines how to build and select the most efficient models for solving binary classification and identification tasks. This process is limited to the series of formal operations, which can be performed by solving any classification task. Efficiency was evaluated with help of ROC curves, and the efficiency of machine learning methods' performance was measured with help of building models' ensembles, using cross-validation, special metrics by training models, and resembling.

Keywords: retail, identification, binary classification, machine learning, frauds, model selection, efficiency increase, models' ensemble, decision making support.

Введение. На сегодняшний день существенно расширяется сфера применения методов поддержки принятия решений и искусственного интеллекта в связи с все возрастающими объемами собираемых данных и появлением новых интеллектуальных алгоритмов и методов машинного обучения. Одним из примеров этих изменений является использование таких устройств, как сканеры товаров в торговле, профессиональная электроника на предприятиях. Сбор данных в современных условиях позволяет наблюдать и оценивать изменения в динамике. Поведение людей может со временем меняться вместе с изменением как их знаний о среде, в которой они находятся, так и вместе с изменениями этой среды (например, поведение мошенников может совершенствоваться в ответ на новые формы контроля), что делает необходимым использования адаптивных методов, для построения которых все большее распространение получают методы машинного обучения.

Наиболее разработанными в настоящее время для решения таких задач являются методы бинарной классификации: линейный дискриминантный анализ [1], машина опорных векторов [2], деревья решений [3], случайный лес [4] и др. Применения этих методов недостаточно для идентификации большого числа возможных ситуаций, однако достаточно для идентификации заведомо негативной ситуации, такой как попытка мошенничества или воровства в магазине самообслуживания на кассе без человека, некорректное использование устройств или ошибки, или выход из строя этих устройств.

Наиболее распространенной проблемой при бинарной классификации данных является то, что наборы данных сильно не сбалансированы. Это означает, что они допускают и неравномерное распределение классов выборок. Кроме того, их распределение сильно пересекается. На примере мошенничества в финансовой сфере и в сфере розничных продаж это означает, что мошенники стараются проводить транзакции таким образом, чтобы они максимально походили на легитимные. Иногда класс легитимных транзакций может превышать класс мошеннических операций в сотни раз.

Идея и область применения существующих методов машинного обучения могут быть различными. Если рассмотреть на примере магазинов самообслуживания, то это могут быть технологии машинного зрения для анализа мимики лиц (см., например, [5]) и контроля за операциями, которые выполняют люди (контроль взвешивания, пронос, см. [6]), технологии обработки данных, собираемых с устройств сканирования, оплаты и т.п., в случае с постоянными клиентами возможно применение методов глубокого обучения для анализа изменений в поведении пользователя (например, в статьях [7], [8] и [9] описывается способ анализа поведения пользователей банка). При работе с выборками для поиска мошенников их количество намного меньше клиентов, не злоупотребляющих доверием. Это привносит свои особенности и требует учета этого факта. Примеры подходов с несбалансированными выборками описаны например в [10] и [11]

Постановка задачи. Рассмотрим постановку задачи на примере выявления мошенников в магазинах, оборудованных технологиями самообслуживания. Важно то, что все данные с устройств самообслуживания (в данном примере – со сканеров) сохраняются, и их можно использовать для того, чтобы обучать и применять модели для предска-

зания поведения мошенников [12]. Стоит иметь в виду то, что мошенничество можно классифицировать как случайное или преднамеренное. Например, покупатель по ошибке может не отсканировать товар либо сделать это преднамеренно.

Ниже в табл. 1 представлены данные, которые могут быть получены из системы самообслуживания (в результате использования переносного сканера) в режиме реального времени.

Таблица 1

Данные, хранимые магазинами, на примере конкурса Data-Mining-Cup 2019 (http://www.data-mining-cup.de)

Название параметра	Описание	Диапазон значений
TrustLevel	Индивидуальный уровень доверия клиента. 6: Высочайшая надежность	{1, 2, 3, 4, 5, 6}
TotalScanTimeInSeconds	Общее время в секундах между первым и последним отсканированным продуктом	Положительное целое число
Grandtotal	Общая стоимость отсканированных продуктов	Положительное десятичное число с 2 знаками после запятой
LineItemVoids	Количество аннулированных сканирований	Целое положительное число
ScansWithoutRegistration	Количество попыток сканирования без какого-либо сканирования (неудачное сканирование)	Целое положительное число или 0
QuantityModification	Число изменений количества товаров для одного из сканируемых продуктов	Целое положительное число или 0
ScannedLineItemsPerSecond	Среднее количество отсканированных продуктов в секунду	Положительное десятичное число
ValuePerSecond	Средняя общая стоимость отсканированных продуктов в секунду	Положительное десятичное число
LineItemVoidsPerPosition	Отношение числа аннулированных сканирований к общему числу аннулированных и неаннулированных сканирований	Положительное десятичное число
Fraud	Классификатор как мошенничество (1) или не мошенничество (0)	{0, 1}

Критерием показателя качества классификации модели выступает следующая функция:

$$R = \sum_{i=1}^{n} \text{score}_i \rightarrow \max,$$

где n – количество наблюдений, score_i – функция, принимающая одно из значений *A*, *B*, *C* или *D* из табл. 2 в зависимости от типа реального класса наблюдения и предсказанного класса.

Таблица 2

Матрица «стоимости» [13]

		Реальные данные	
		Легитимная транзакция	**Мошенническая транзакция**
Модельные значения	**Легитимная транзакция**	**A**	**– B**
	Мошенническая транзакция	**– C**	**D**

Интерпретация матрицы стоимости выглядит следующим образом: ритэйлер получает прибыль в размере **D** у.е. за каждую правильно выявленную попытку мошенничества. Тем не менее на каждый случай мошенничества, который классифицируется системой как легитимная продажа, продавец теряет **B** у.е. Клиент, ложно обвиненный в мошенничестве, может не вернуться в этот магазин, что составляет **C** у.е. в убыток для ритэйлера.

Для имитации реальной ситуации расчёты будем проводить на двух наборах данных. Одним является размеченный набор данных (выбранный нами этот набор данных 1879 наблюдений), на котором мы обучаем и проверяем модель. Второй набор данных является контрольным и содержит 489 121 наблюдение. Для использования первого набора данных его разбиваем на две части методом стратифицированного отбора [14, 15] для сохранения соотношения классов (мошенников и добропорядочных покупателей) в выборках. В результате работу проводим с тремя наборами данных: обучающим (940 наблюдений), тестовым (939 наблюдений), контрольным (489 121 наблюдений).

Методология решения поставленной задачи сводится к выбору наилучшего метода, позволяющего учитывать специфику данных.

Первым шагом является прямое использование известных методов классификации на обучающей выборке [16]. Далее можно повысить качество моделей путем использования изменений в процессе их обучения: 1) кросс-валидации [17]; 2) введение специализированной метрики; 3) использование гиперпараметризации моделей (задание диапазонов значений параметров оказывающих наибольшее влияние); 4) использование ресемплинга вместо кросс-валидации [18].

Этап 1. Подготовка данных.

Шаг 1.0. Очистка данных.

Шаг 1.1. Добавление вычислимых показателей в имеющийся набор данных.

Шаг 1.2. Уменьшение количества показателей в наборах данных путём исключения показателей с околонулевой дисперсией, коррелирующих показателей и показателей, являющихся линейной комбинацией.

Шаг 1.3. Разбиение размеченного набора данных путём стратифицированного отбора для сохранения соотношения классов (получаем три набора данных: обучающий, тестовый, контрольный).

Этап 2. Выбор метода.

Шаг 2.1. Определение множества методов машинного обучения для проверки их эффективности.

Шаг 2.2. Выбор наилучших методов для проверки возможностей повышения качества их работы.

Этап 3. Повышение эффективности единичных методов.

Шаг 3.1. Обучение отобранных методов с использованием кросс-валидации без метрики качества, без настройки гиперпараметров и проверка результатов на тестовой и контрольной выборке.

Шаг 3.2. Обучение отобранных методов с использованием кросс-валидации, сметрикой качества, без настройки гиперпараметров и проверка результатов на тестовой и контрольной выборке.

Шаг 3.3. Обучение отобранных методов с использованием кросс-валидации, с метрикой качества, с настройкой гиперпараметров и проверка результатов на тестовой и контрольной выборке.

Шаг 3.4. Обучение отобранных методов с использованием ресемплинга вместо кросс-валидации, с метрикой качества, с настройкой гиперпараметров и проверка результатов на тестовой и контрольной выборке.

Этап 4. Объединение моделей (ансамбли или композиции моделей).

Шаг 4.1. Выбор моделей с наибольшим значением метрики на тестовых данных и наибольшей разницей площадей под ROC-кривыми.

Шаг 4.2. Получение откликов выбранных моделей на тестовой выборке и объединение полученных откликов в один набор данных ответа (создание ансамблей моделей разными методами объединения).

Шаг 4.3. Подсчёт результатов для разных способов объединения и выбор наилучшего алгоритма объединения.

Рис. 1. Алгоритм выбора и улучшения для решения задачи бинарной классификации [19]

Повышение качества предсказания может произойти не только из-за изменения моделей, но и из-за обучающих данных: 1) введение в обучающую выборку новых вычисляемых показателей, исходя из знаний о предметной области; 2) исключение показателей с околонулевой дисперсией, коррелирующих показателей и показателей, являющихся линейной комбинацией.

Дальнейшее улучшение возможно при использовании ансамблей или композиций моделей. Их построение возможно на разных принципах. Поэтому необходимо будет выбрать лучший метод объединения моделей (наиболее часто используемыми являются методы Bagging-a, Busting-a и Stacking-a). При этом эффект от объединения моделей можно получить, только если разные методы определяют разные ситуации. В результате работа алгоритма выбора и адаптации метода классификации для конкретной задачи описывается алгоритмом, приведенным на рис. 1.

Решение. Для проверки работы алгоритма будем использовать 4 модели: Support-VectorMachine (метод опорных векторов), ClassificationandRegressionTree (деревья решений), NeuralNet (однослойный персептрон) и метод kNN (k-ближайших соседей).

Для решения поставленной задачи сгенерируем из исходных данных два новых набора данных. Первый набор данных является исходным. Во втором наборе данных появляются новые параметры, рассчитанные на основе исходных данных. Рассчитанные признаки приведены в табл. 3.

Таблица 3

Вычисляемые параметры при решении задачи поиска мошенников в ритэйле

Название параметра	Описание поля	Формула из параметров табл. 1
ne_otm	Количество неотменённых заказов	totalScanTimeInSeconds * scannedLineItemsPerSecond
otm_i_ne_otm	Отношение количества аннулированных сканирований к неаннулированным	lineItemVoids + ne_otm
sec_na_1_udach_scan	Отношение общего времени к количеству удачных сканирований	totalScanTimeInSeconds / otm_i_ne_otm
udach_i_neudach_scan	Количество удачных и неудачных сканирований	otm_i_ne_otm + scansWithoutRegistration
dolya_neudach_scan	Отношение количества удачных сканирований к неудачным	scansWithoutRegistration / udach_i_neudach_scan
sec_na_1_scan	Отношение общего времени в магазине к общему количеству сканирований	totalScanTimeInSeconds / udach_i_neudach_scan

Второй набор данных подвергаем процедуре отбора признаков, в результате получаем третий набор данных. Для поиска незначимых параметров выполним три операции. Первая – выявление признаков с околонулевой дисперсией. Вторая – выявление признаков, которые имеют коэффициент корреляции выше 0,75. Третья – поиск параметров, являющихся линейной комбинацией других. На используемых данных первая и третья проверки не выявили неинформативных признаков. По результатам второй проверки были найдены три неинформативных предиктора, которые приведены в табл. 4.

Таблица 4

Исключенные из второго набора данных неинформативные параметры

Название параметра	Описание поля	Формула из параметров табл. 1
otm_i_ne_otm	Отношение количества аннулированных сканирований к неаннулированным	lineItemVoids + ne_otm
udach_i_neudach_scan	Количество удачных и неудачных сканирований	otm_i_ne_otm + scansWithoutRegistration
sec_na_1_scan	Отношение общего времени в магазине к общему количеству сканирований	totalScanTimeInSeconds / udach_i_neudach_scan

После генерации новых параметров и исключения неинформативных получено три набора данных с различным числом предикторов: 1) 9 предикторов и 1 метка класса (исходные данные), 2) 15 предикторов и 1 метка класса (после генерации новых параметров), 3) 12 предикторов и 1 метка класса (после исключения неинформативных предикторов).

Для оценки эффективности будем использовать метрику, описанную матрицей стоимости (см. табл. 2). Для данной задачи выбраны коэффициенты A – 5, B – 5, C – 25, D – 0. С такими параметрами метрика дает больший штраф за ошибку первого рода, что логично для задачи поиска мошеннических транзакций.

Для проверки работы алгоритмов разбиваем размеченные данные путём стратифицированного отбора для сохранения соотношения классов. Получили три набора данных: обучающий, тестовый, контрольный (неразмеченные данные).

Обучим модели на обучающих выборках и выполним с первого по третий этапы, описанные в алгоритме (см. рис. 1).

Результаты работы на тестовой выборке приведены на рис. 2. Полученные графики сгруппированы по типу моделей и операциям повышения эффективности. Внутри каждого графика отложены значения метрики для каждого набора данных с различным числом предикторов.

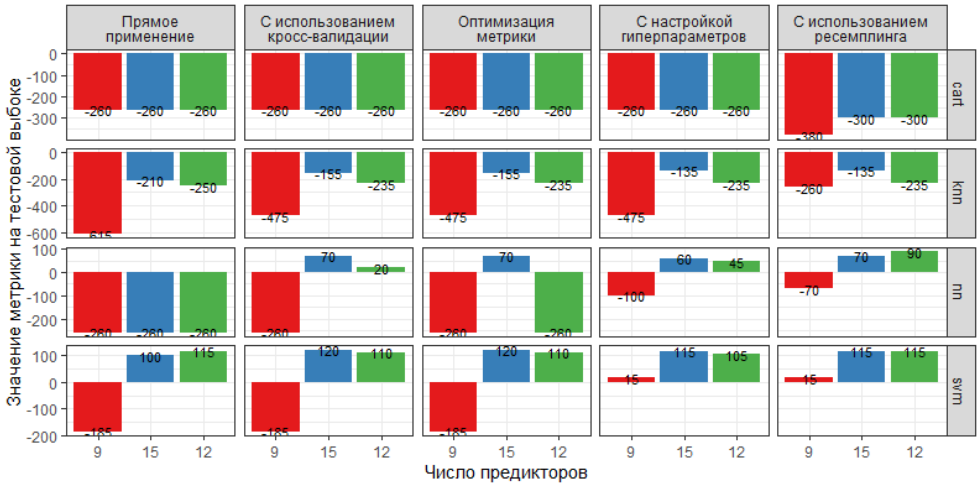

Рис. 2. Результаты оценки метрики всех алгоритмов на тестовой выборке (красный – набор данных с 9 предикторами, синий – набор данных с 15 предикторами, зеленый – набор данных с 12 предикторами)

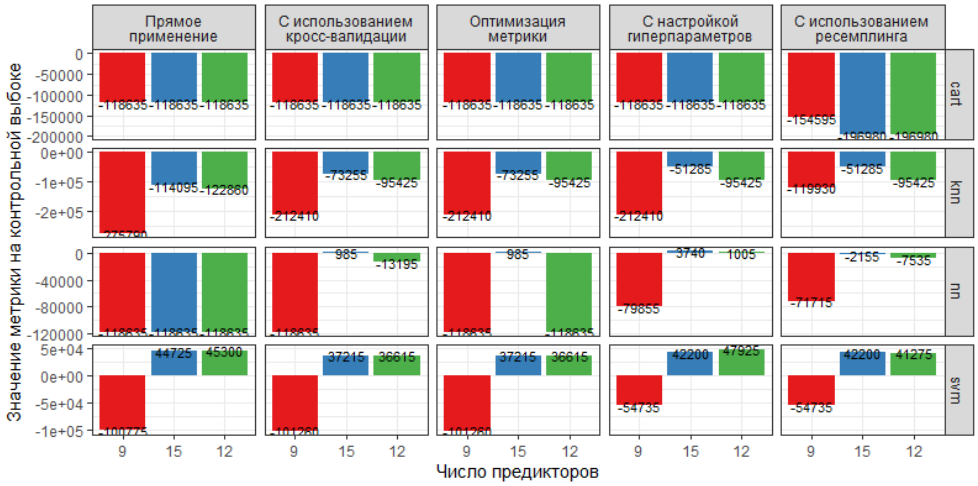

Рис. 3. Результаты оценки метрики всех алгоритмов на контрольной выборке

На рис. 3 изображены результаты работы на контрольной выборке. Полученные результаты отличаются от результатов тестовой выборки, но видно, что тенденции повышения эффективности сохраняются.

Это свидетельствует о том, что использование приведенного на рис. 1 алгоритма позволяет решать задачу выбора операций, приводящих к выбору методов для решения задач классификации и операций, приводящих к их наибольшему повышению эффективности.

Можно предположить, что разные модели могут идентифицировать различные ситуации, в том числе делать отличающиеся по логике ошибки (чем больше площадь, не являющаяся пересечением ROC-кривых (рис. 4) моделей с близкими значениями метрики, тем более разные случаи они умеют находить). Это означает, что объединение моделей может дать повышение количества верно идентифицированных ситуаций за счет увеличения количества верно определенных мошенников и снижения количества ошибок.

Для выбора моделей для ассемблирования будем опираться на оценку результатов на тестовой выборке (выбирать модели с наибольшими значениями метрики). В результате согласно рис. 2 выберем методы: машину опорных векторов (svm) на выборках с 15 и 12; нейронную сеть (nn) на наборе с 15 предикторами с использованием кросс-валидации, с оптимизацией метрики, с настройкой гиперпараметров, с использованием ресемплинга и на наборе с 12 предикторами с использованием ресемплинга.

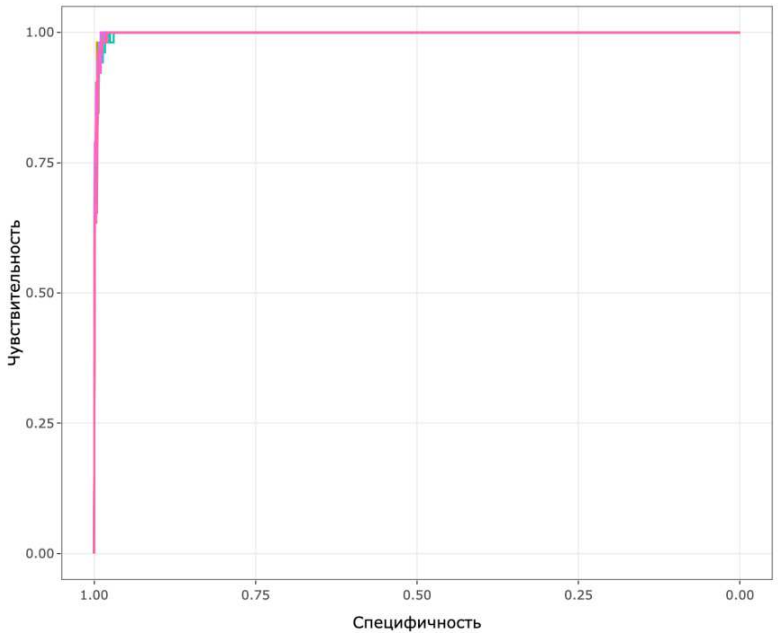

Рис. 4. ROC-кривые для моделей, выбранных для ассамблирования

Проверим эффект от объединения моделей [20] на основе двух способов объединения: 1) голосование моделей на основе значений классов и 2) на основе вычисления суммы вероятностей для класса по результатам оценки вероятностей для классов с отдельными моделями.

Работа алгоритма голосования на основе значений классов сводится к суммированию значений классов и определению границы, после превышения которой наблюдению присваивается одно значение (например, 1), а иначе – другое (например, 0), как приведено на рис. 5.

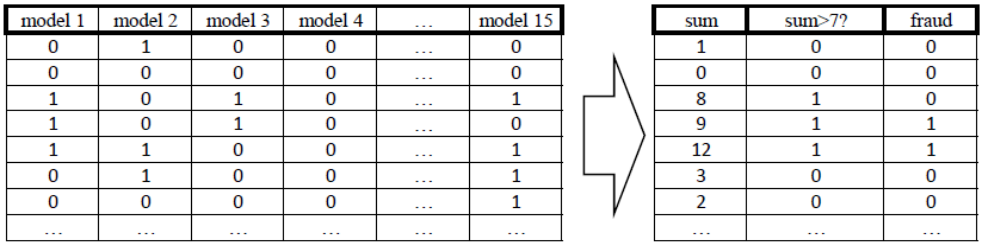

Рис. 5. Пример работы алгоритма объединения моделей на основе значений класса

Эффективность работы алгоритма будет зависеть от выбранного порогового значения. В табл. 5 приведено количество наблюдений, соответствующих классу 1 при разных пороговых значениях на тестовой выборке. Тогда выбор порогового значения становится аргументом задачи максимизации ($\underset{\text{sum}}{\text{argmax}}$(метрика)).

Таблица 5

Значение метрики для тестовой выборки для разного количества положительных откликов моделей

sum>0	**sum>1**	**sum>2**	**sum>3**	**sum>4**	**sum>5**	**sum>6**	**sum>7**	**sum>8**	**sum>9**	**sum>10**	**sum>11**	**sum>12**	**sum>13**	**sum>14**
65	65	65	65	55	120	120	110	100	125	125	130	120	120	100

На рассмотренной выборке наибольшее количество совпадений классов с тестовыми значениями наблюдается при пороговом значении, равном 11. При объединении моделей на основе использования вероятностей вычисление конечной вероятности может осуществиться на основе формулы суммирования вероятностей:

$$\big(P(A+B) = P(A) + P(B) - P(AB)\big).$$

Суммирование моделей будем осуществлять в порядке уменьшения значения вероятности метрики на тестовой выборке (пример работы алгоритма на тестовой выборке приведен на рис. 6). Выбор порогового значения вероятности осуществляется так же, как и в предыдущем случае ($\arg\max_{0 \le P \le 1}$(метрика)).

model 1	model 2	model 3	model 4	…	model 15
0.0001	0.9941	0.0002	0.0001	…	0.0001
0.0023	0.0011	0.0019	0.0001	…	0.0231
0.9563	0.0001	0.8216	0.0001	…	0.9999
0.9648	0.0013	0.9245	0.0001	…	0.0001
0.8976	0.7489	0.0034	0.0001	…	0.8952
0.0002	0.8988	0.0077	0.0001	…	0.9112
0.0016	0.0014	0.0358	0.0001	…	0.9655
…	…	…	…	…	…

prob sum	prob sum>0.8?	fraud
0.1541	0	0
0.6587	0	0
0.6411	0	0
0.8869	1	1
0.9746	1	1
0.3651	0	0
0.4321	0	0
…	…	…

Рис. 6. Пример работы алгоритма объединения моделей на основе вероятностей принадлежности к 1-му классу

Для проверки работоспособности модели выберем пороговое значение вероятностиравное 0,987 (при превышении этого значения наблюдению присваивается 1-й класс, иначе класс 0).

Полученные после объединения указанными выше способами модели приведены на рис. 7. На рисунке видно, что объединение моделей может приводить не только к повышению эффективности, но и к ее снижению. Таким образом, выбор метода ансамблирования является задачей сопоставительного выбора. В том числе в зависимости задачи могут появляться новые способы объединения, опирающиеся на закономерности рассматриваемой предметной области.

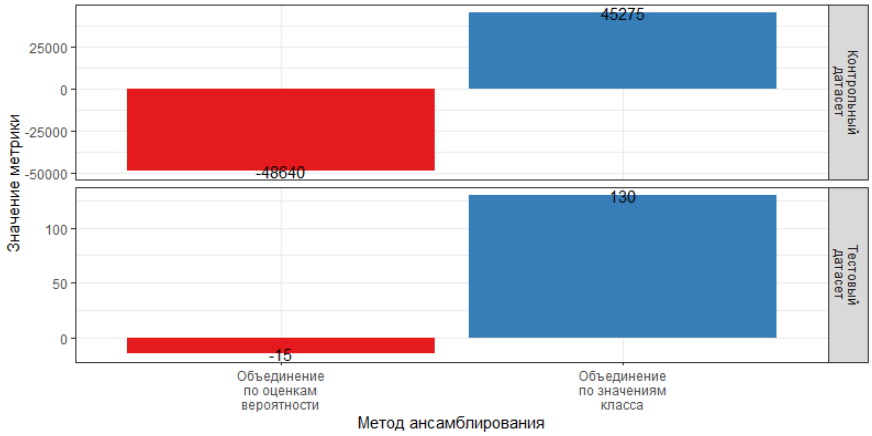

Рис. 7. Результаты оценки метрики двух методов ансамблирования для тестовой и контрольной выборок (красный – суммирование вероятностей 1-го класса, синий – суммирование значений классов)

Дискуссия. Полученные результаты показывают, что задача построения алгоритма идентификации является многоэтапной, многосвязной, инвариантной, зависящей от используемых статистических данных. Например, может быть изменена пропорция разбиения данных на тестовую и обучающую выборки, для объединения моделей можно делать множественные разбиения, когда различные данные будут попадать в тестовые и обучающие наборы данных. Есть модели, которые описанные в статье подходы используют в процессе самообучения (такая модель, как случайный лес, использует ресемплинг, настройку гиперпараметров и отбор признаков).

В процессе использования полученной модели могут меняться типы поведения клиентов, поэтому модель потребует переобучения и адаптации к новым условиям. Это приводит к постановке задачи выбора окна времени в прошлое, которое необходимо будет постоянно использовать для переобучения модели и получения баланса между ее чрезмерной инертностью и чувствительностью к изменениям [21].

Закономерно, что, используя модель, мы лишь предполагаем, что некоторые клиенты являются добропорядочными, а некоторые – мошенниками. Скорее всего, объем выборок с предположениями и точными знаниями (которые были получены в результате досмотра или наблюдения) сильно отличаются. Это позволяет говорить о том, что в процессе работы следует использовать обучение с подкреплением, а значит, сам процесс обучения моделей тоже становится многоэтапным инвариантным процессом, при котором использование размеченных и неразмеченных (не подтвержденных) данных может быть различным.

Выводы. Описанный в статье алгоритм позволяет получать, находить и строить эффективный метод решения задачи бинарной классификации с учетом особенностей имеющихся статистических данных и задачи, решение которой необходимо.

Применительно к практической деятельности описанное в статье представление задач управления позволяет конструировать эмпирические модели исследуемой предметной области, позволяющие улавливать закономерности, связанные с их спецификой, что не всегда возможно при построении математических моделей, основанных на использовании

известных физических, экономических и других закономерностей. При этом используемый подход к моделированию процессов поддержки принятия решений позволяет рассматривать процессы, происходящие на любом уровне управления, независимо от типов и видов управления.

Библиографический список

1. Прикладная статистика: Классификация и снижение размерности / С.А. Айвазян [et al.]. – М., 1989. – 607 с.

2. Вапник В.Н., Червоненкис А.Я. Теория распознавания образов. – М.: Наука, 1974. – 487 с.

3. Classification and Regression Trees / L. Breiman [et al.]. – Belmont (CA): Wadsworth Int. Group, 1984. – 368 p.

4. Breiman L. Random forests // Mach. Learn. – 2001. – Vol. 45, № 1. – P. 5–32.

5. Zhang L., Wang Z. A multi-view camera-based anti-fraud system and its applications // J. Vis. Commun. Image Represent. – 2018. – Vol. 55. – P. 263–269.

6. A pattern discovery approach to retail fraud detection / P. Gabbur [et al.] // Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining – KDD ’11. – New York, USA: ACM Press, 2011. – P. 307.

7. Combining unsupervised and supervised learning in credit card fraud detection / F. Carcillo [et al.] // Inf. Sci. (Ny). – 2019. – Vol. S002002551.

8. Kim E. Champion-challenger analysis for credit card fraud detection: Hybrid ensemble and deep learning // Expert Syst. Appl. – 2019. – Vol. 128. – P. 214–224.

9. Chouiekh A., EL Haj E.H.I. ConvNets for Fraud Detection analysis // Procedia Comput. Sci. – 2018. – Vol. 127. – P. 133–138.

10. Sridhar S., Karthigayani P. A novel approach for decision tree occlusion on detection (DTOD) classifier for face verification and estimation of age using back propagation Neural Network (BPNN) // J. Computer Sci. Inf. Technol. Res. – 2013. – Vol. 3, № 1. – P. 1–10.

11. Ahmed I., Pariente A., Tubert-Bitter P. Class-imbalanced subsampling lasso algorithm for discovering adverse drug reactions // Stat. Methods Med. Res. – 2018. – Vol. 27, № 3. – P. 785–797.

12. Demirci Orel F., Kara A. Supermarket self-checkout service quality, customer satisfaction, and loyalty: Empirical evidence from an emerging market // J. Retail. Consum. Serv. – 2014. – Vol. 21, № 2. – P. 118–129.

13. Мыльников Л.А., Колчанов С.А. Методика выявления ключевых параметров инновационных проектов на основе статистических данных // Экономический анализ теория и практика. – 2012. – № 5(260). – С. 22–28.

14. Shahrokh Esfahani M., Dougherty E.R. Effect of separate sampling on classification accuracy // Bioinformatics. – 2013. – Vol. 30, № 2. – P. 242–250.

15. An Introduction to Statistical Learning: with Applications in R / G. James [et al.]. – Springer New York, 2014.

16. Мыльников Л.А. Статистические методы интеллектуального анализа данных. – Пермь: Изд-во Перм. нац. исслед. политехн. ун-та, 2018. – 168 с.

17. Kohavi R. A Study of Cross-validation and Bootstrap for Accuracy Estimation and Model Selection // Proceedings of the 14th International Joint Conference on Artificial Intelligence. Vol. 2. – San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1995. – P. 1137–1143.

18. Шитиков В.К., Розенберг Г.С. Рандомизация и бутстреп: статистический анализ в биологии и экологии с использованием R. – Тольятти: Кассандра, 2013. – 314 с.

19. Интеллектуальный анализ данных в управлении производственными системами: подходы и методы / Л.А. Мыльников [et al.]. – М.: Библиоглобус, 2017. – 332 с.

20. Wolpert D.H. Stacked generalization // Neural Networks. – 1992. – Vol. 5, № 2. – P. 241–259.

21. Mylnikov L.A., Kulikov M.V., Krause B. The selection of optimal control of the operation modes of heterogeneous duplicating equipment based on statistical models with learning // Int. J. Mech. Eng. Technol. – 2018. – Vol. 9, № 9.

References

1. Aivazian S.A. et al. Prikladnaia statistika: Klassifikatsiia i snizhenie razmernosti [Applied statistics: classification and reduction of dimensions]. Moscow, 1989. 607 p.

2. Vapnik V.N., Chervonenkis A.Ia. Teoriia raspoznavaniia obrazov [Pattern Recognition Theory]. Moscow: Nauka, 1974. 487 p.

3. L. Breiman et al. Classification and Regression Trees. Belmont (CA): Wadsworth Int. Group, 1984. 368 p.

4. Breiman L. Random forests. *Mach. Learn*, 2001, vol. 45, no. 1, pp. 5-32.

5. Zhang L., Wang Z. A multi-view camera-based anti-fraud system and its applications. *J. Vis. Commun. Image Represent*, 2018, vol. 55, pp. 263-269.

6. Gabbur P. et al. A pattern discovery approach to retail fraud detection. *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '11*. New York, USA: ACM Press, 2011, P. 307.

7. Carcillo F. et al. Combining unsupervised and supervised learning in credit card fraud detection. *Inf. Sci. (Ny)*, 2019, vol. S002002551.

8. Kim E. Champion-challenger analysis for credit card fraud detection: Hybrid ensemble and deep learning. E*xpert Syst. Appl*, 2019, vol. 128, pp. 214-224.

9. Chouiekh A., EL Haj E.H.I. ConvNets for Fraud Detection analysis. *Procedia Comput. Sci*, 2018, vol. 127, pp. 133-138.

10. Sridhar S., Karthigayani P. A novel approach for decision tree occlusion on detection (DTOD) classifier for face verification and estimation of age using back propagation Neural Network (BPNN). *J. Computer Sci. Inf. Technol. Res*, 2013, vol. 3, no. 1, pp. 1-10.

11. Ahmed I., Pariente A., Tubert-Bitter P. Class-imbalanced subsampling lasso algorithm for discovering adverse drug reactions. *Stat. Methods Med. Res*, 2018, vol. 27, no. 3, pp. 785-797.

12. Demirci Orel F., Kara A. Supermarket self-checkout service quality, customer satisfaction, and loyalty: Empirical evidence from an emerging market. *J. Retail. Consum. Serv*, 2014, vol. 21, no. 2, pp. 118-129.

13. Myl'nikov L.A., Kolchanov S.A. Metodika vyiavleniia kliuchevykh parametrov innovatsionnykh proektov na osnove statisticheskikh dannykh [Methodology for identifying key parameters of innovative projects based on statistical data]. *Ekonomicheskii analiz teoriia i praktika*, 2012, no. 5(260), pp. 22-28.

14. Shahrokh Esfahani M., Dougherty E.R. Effect of separate sampling on classification accuracy. *Bioinformatics*, 2013, vol. 30, no. 2, pp. 242-250.

15. James G. et al. An Introduction to Statistical Learning: with Applications in R. Springer New York, 2014.

16. Myl'nikov L.A. Statisticheskie metody intellektual'nogo analiza dannykh [Statistical methods of intelligent data analysis]. Perm': Permskii natsional'nyi issledovatel'skii politekhnicheskii universitet, 2018. 168 p.

17. Kohavi R. A Study of Cross-validation and Bootstrap for Accuracy Estimation and Model Selection. *Proceedings of the 14th International Joint Conference on Artificial Intelligence*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1995, vol. 2, pp. 1137-1143.

18. Shitikov V.K., Rozenberg G.S. Randomizatsiia i butstrep: statisticheskii analiz v biologii i ekologii s ispol'zovaniem R [Randomization and bootstrap: statistical analysis in biology and ecology using R]. Tol'iatti: Kassandra, 2013. 314 p.

19. Myl'nikov L.A. et al. Intellektual'nyi analiz dannykh v upravlenii proizvodstvennymi sistemami: podkhody i metody [Intelligent data analysis in the management of production systems (approaches and methods)]. Moscow: Biblio-globus, 2017. 332 p.

20. Wolpert D.H. Stacked generalization. *Neural Networks*, 1992, vol. 5, no. 2, pp. 241-259.

21. Mylnikov L.A., Kulikov M.V., Krause B. The selection of optimal control of the operation modes of heterogeneous duplicating equipment based on statistical models with learning. *Int. J. Mech. Eng. Technol*, 2018, vol. 9, no. 9.

Сведения об авторах

Мыльников Леонид Александрович (Пермь, Россия) – кандидат технических наук, доцент кафедры «Микропроцессорные средства автоматизации» Пермского национального исследовательского политехнического университета (614990, Пермь, Комсомольский пр., 29, e-mail: leonid.mylnikov@pstu.ru).

Морозов Алексей Сергеевич (Пермь, Россия) – магистрант кафедры «Микропроцессорные средства автоматизации» Пермского национального исследовательского политехнического университета (614990, Пермь, Комсомольский пр., 29, e-mail: morozov.leha@mail.ru).

Пухарева Дарья Вадимовна (Пермь, Россия) – магистрант кафедры «Информационные технологии и автоматизированные системы» Пермского национального исследовательского политехнического университета (614990, Пермь, Комсомольский пр., 29, e-mail: dasha.pukhareva@yandex.ru).

About the authors

Mylnikov Leonid Aleksandrovich (Perm, Russian Federation) is a Ph.D. in Technical Science, Associated Professor of Microprocessor Automation Means Department Perm National Research Polytechnic University (614990, Perm, 29, Komsomolskypr., e-mail: leonid.mylnikov@pstu.ru).

Morozov Aleksey Sergeevich (Perm, Russian Federation) is a Master Student at Microprocessor Automation Means Department Perm National Research Polytechnic University (614990, Perm, 29, Komsomolskypr., e-mail: morozov.leha@mail.ru).

Pukhareva Daria Vadimovna (Perm, Russian Federation) is a Master Student atInformation Technology and Automation Systems Department Perm National Research Polytechnic University (614990, Perm, 29, Komsomolskypr., e-mail: dasha.pukhareva@yandex.ru).

Получено 27.01.2020