МОСКОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
имени М.В. ЛОМОНОСОВА

На правах рукописи

Ракитько Александр Сергеевич

Идентификация значимых факторов с помощью функционала ошибки

1.1.4 — теория вероятностей и математическая статистика

АВТОРЕФЕРАТ
диссертации на соискание ученой степени
кандидата физико-математических наук

Москва — 2023

Диссертация подготовлена на кафедре теории вероятностей механико-математического факультета МГУ имени М.В.Ломоносова.

Научный руководитель: **Булинский Александр Вадимович**, доктор физико-математических наук, профессор

Официальные оппоненты: **Ульянов Владимир Васильевич,** доктор физико-математических наук, профессор, МГУ имени М.В.Ломоносова, профессор кафедры математической статистики факультета ВМК

Шоргин Сергей Яковлевич, доктор физико-математических наук, профессор, ФИЦ ИУ РАН, главный научный сотрудник отдела информационных технологий управления и моделирования информационных систем

Кожевин Алексей Александрович, кандидат физико-математических наук, ООО «Газпромнефть-ЦР», эксперт по машинному обучению

Защита диссертации состоится «28» июня 2023 г. в 15 часов 00 минут на заседании диссертационного совета МГУ.011.3 Московского государственного университета имени М.В.Ломоносова по адресу: Российская Федерация, 119991, ГСП-1, Москва, Ленинские горы, д.1, МГУ, механико-математический факультет, аудитория 16-10.

E-mail: mexmat_disser85@mail.ru

С диссертацией можно ознакомиться в отделе диссертаций научной библиотеки МГУ имени М.В.Ломоносова (Ломоносовский просп., д. 27) и на портале: https://dissovet.msu.ru/dissertation/011.3/2520.

Автореферат разослан «27» мая 2023 года.

Заместитель председателя
диссертационного совета,
доктор физико-математических наук,
доцент И.С. Ломов

Ученый секретарь
диссертационного совета,
доктор физико-математических наук,
доцент Н.А. Раутиан

Общая характеристика работы

Актуальность темы. В последние годы благодаря развитию информационных технологий наблюдается значительный рост объема данных, доступных для анализа. В связи с этим, огромное внимание уделяется современным исследовательским областям, связанным с анализом больших массивов данных, которые в англоязычной литературе носят названия Data Science, Data mining, Big Data, Machine Learning, Deep Learning[1,2]. Это объясняется тем, что увеличение количества анализируемых данных предоставило возможность обнаруживать более сложные зависимости между переменными, нежели, например, линейные. Ярким примером является возрастающая популярность алгоритмов, в основе которых лежит архитектура нейронных сетей[3,4].

Одной из областей, в которых исследователи неизбежно сталкиваются с необходимостью анализа данных высоких размерностей (больших данных), является биоинформатика. Прорыв в данной области был обусловлен прогрессом технологий расшифровки генома человека (Next-Generation Sequencing – секвенирование следующего поколения[5]). Развивающиеся информационные технологии, в том числе и квантовые вычисления, в перспективе позволят еще снизить стоимость и ускорить анализ генетических данных[6]. Задача выявления факторов, ассоциированных с риском возникновения некоторого заболевания, является одной из наиболее частых в современной биостатистике[7]. В некоторых исследованиях с помощью статистических тестов проверяется гипотеза о зависимости между генами-кандидатами, предположительно связанными с болезнью, и наличием заболевания у пациента[8]. В других исследованиях осуществляется полногеномный поиск ассоциаций, известный в англоязычной литературе как

---

[1]Murphy K. P. Probabilistic machine learning: an introduction. — MIT press, 2022. – 864 p.

[2]Грушо А. А., Грушо Н. А., Забежайло М. И., Смирнов Д. В. [и др.]. Поиск аномалий в больших данных // Системы и средства информатики. — 2022. — Т. 32, № 1. — С. 160—167.

[3]Roberts D. A., Yaida S., Hanin B. The principles of deep learning theory. — Cambridge University Press Cambridge, MA, USA, 2022. — 472 p.

[4]Бобрикова Е. В., Платонова А. А., Гайдамака Ю. В., Шоргин С. Я. Пример применения аппарата нейронных сетей при назначении модуляционно-кодовой схемы планировщиком базовой станции сети 5G // Системы и средства информатики. — 2021. — Т. 31, № 3. — С. 135—143.

[5]Ребриков Д. В., Коростин Д. О., Шубина Е. С., Ильинский В. В. NGS: высокопроизводительное секвенирование. — БИНОМ. Лаборатория знаний, 2015. — 232 с.

[6]Boev A., Rakitko A., Usmanov S., Kobzeva A., [et al.]. Genome assembly using quantum and quantum-inspired annealing // Scientific Reports. — 2021. — Vol. 11, no. 1. — P. 13183.

[7]Mills M. C., Barban N., Tropf F. C. An introduction to statistical genetic data analysis. — MIT Press, 2020. — 432 p.

[8]Berseneva A., Kovalenko E., Vergasova E., Prohorov A., [et al.]. Association of common genetic variants with body mass index in Russian population // European Journal of Clinical Nutrition. — 2023.

GWAS (Genome-wide association studies)[9,10,11,12]. В таких исследованиях рассматривается зависимость между некоторой случайной функцией Y, обозначающей фенотип (наличие или отсутствие болезни, биохимический показатель крови, способность к обучению и т.д.) и генетическими факторами $X_1, \ldots, X_n$, $n \in \mathbb{N}$, находящимися почти во всех генах и межгенных пространствах. В современных исследованиях количество факторов n может достигать нескольких миллионов. Например, в геноме человека есть порядка 10 миллионов позиций, мутации в которых встречаются достаточно частно и могут быть связаны с различными заболеваниями. В последние годы активно разрабатывается математический аппарат анализа данных высоких размерностей, например, генетических, когда число факторов n, описывающих одного индивидуума, соизмеримо с размером выборки N[13]. При этом по-настоящему связанными с откликом являются лишь некоторые из факторов $X_1, \ldots, X_n$, $n \in \mathbb{N}$. Для практических медицинских задач крайне важно знать, какие именно факторы влияют на функцию отклика. Это позволяет строить прогностические модели для предсказания риска развития заболевания[14], понимать патогенез заболеваний на молекулярно-генетическом уровне и находить мишени для лекарств. Возможно, в будущем в практику войдут и технологии редактирования генома.

Во многих случаях применяется двухэтапная процедура поиска значимых факторов. На первом этапе проводится однофактортный анализ, например, тест хи-квадрат Пирсона. По результатам теста отбираются факторы, показавшие наибольшую зависимость с изучаемой характеристикой. На втором этапе применяется многофакторный анализ понижения размерности данных наблюдений. В последние десятилетия активно разрабатывались такие методы как логистическая регрессия, случайные леса,

[9]COVID-19 Host Genetics Initiative. Mapping the human genetic architecture of COVID-19 // Nature. — 2021. — Vol. 600, no. 7889. — P. 472—477.

[10]Kasyanov E., Rakitko A., Rukavishnikov G., Golimbet V., [et al.]. Contemporary Genome-Wide Association Studies in Depression: The Critical Role of Phenotyping // Neuroscience and Behavioral Physiology. — 2022. — Vol. 52, no. 6. — P. 826—835.

[11]Pinakhina D., Yermakovich D., Vergasova E., Kasyanov E., [et al.]. GWAS of depression in 4,520 individuals from the Russian population highlights the role of MAGI2 (S-SCAM) in the gut-brain axis // Frontiers in Genetics. — 2023. — Vol. 13. — P. 3571.

[12]Kibitov A., Rakitko A., Kasyanov E., Yermakovich D., [et al.]. Genome-wide association study of depression symptoms using online self-questionnaires in the Russian population cohort: preliminary results // European Psychiatry. — 2022. — Vol. 65, S1. — S327—S327.

[13]Fujikoshi Y., Ulyanov V. V. Non-asymptotic Analysis of Approximations for Multivariate Statistics. — Springer Singapore, 2020. — 130 p.

[14]Borisevich D., Schnurr T. M., Engelbrechtsen L., Rakitko A., [et al.]. Non-linear interaction between physical activity and polygenic risk score of body mass index in Danish and Russian populations // Plos One. — 2021. — Vol. 16, no. 10. — e0258748.

LASSO и байесовские методы[15], условная энтропия Шеннона[16,17,18], комбинации упомянутых методов[19] и другие.

Как уже отмечалось выше, для некоторых болезней характерны ситуации, когда по отдельности факторы могут давать незначительный вклад в развитие заболевания. Однако, их определенные комбинации могут приводить к существенному увеличению риска болезни. С практической точки зрения это означает необходимость применения нелинейных моделей, которые уже используются в различных областях: от биостатистики[20] до усвоения данных (оценка состояния системы на основании текущих наблюдений, исторических наблюдений и модельных предположений) в гидрометеорологии[21,22]. С целью выявления комбинаций факторов, влияющих на риск болезни, был предложен метод MDR (multifactor dimensionality reduction)[23]. Сейчас этот алгоритм активно используется в практических исследованиях с целью выявления эффекта взаимодействия генов для различных заболеваний[24]. Были разработаны различные модификации данного метода, см., например, обзор[25]. В последние годы продолжают появляться работы, посвященные улучшениям и модификациям MDR алгоритма. Ф. Абегаз и коллеги исследовали три варианта MDR

[15]Pudjihartono N., Fadason T., Kempa-Liehr A. W., O'Sullivan J. M. A review of feature selection methods for machine learning-based disease risk prediction // Frontiers in Bioinformatics. — 2022. — Vol. 2. — P. 927312.

[16]Bulinski A., Kozhevin A. Statistical estimation of conditional Shannon entropy // ESAIM: Probability and Statistics. — 2019. — Vol. 23. — P. 350—386.

[17]Bulinski A., Kozhevin A. Statistical estimation of mutual information for mixed model // Methodology and Computing in Applied Probability. — 2021. — Vol. 23. — P. 123—142

[18]Kozhevin A. A. Feature selection based on statistical estimation of mutual information // Siberian Electronic Mathematical Reports. — 2021. — Vol. 18, no. 1. — P. 720—728.

[19]Guo H., Yu Z., An J., Han G., [et al.]. A two-stage mutual information based Bayesian Lasso algorithm for multi-locus genome-wide association studies // Entropy. — 2020. — Vol. 22, no. 3. — P. 329.

[20]См. сноски 7 и 14 выше.

[21]Tsyrulnikov M., Rakitko A. A hierarchical Bayes ensemble Kalman filter // Physica D: Nonlinear Phenomena. — 2017. — Vol. 338. — P. 1—16.

[22]Tsyrulnikov M., Rakitko A. Impact of non-stationarity on hybrid ensemble filters: A study with a doubly stochastic advection-diffusion-decay model // Quarterly Journal of the Royal Meteorological Society. — 2019. — Vol. 145, no. 722. — P. 2255—2271.

[23]Ritchie M. D., Hahn L. W., Roodi N., Bailey L. R., [et al.]. Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer // The American Journal of Human Genetics. — 2001. — Vol. 69, no. 1. — P. 138—147.

[24]Zhang C., Qin Q., Li Y., Zheng X., [et al.]. Multifactor dimensionality reduction reveals the effect of interaction between ERAP1 and IFIH1 polymorphisms in psoriasis susceptibility genes // Frontiers in Genetics. — 2022. — Vol. 13.

[25]Gola D., Mahachie John J. M., Steen K. van, König I. R. A roadmap to multifactor dimensionality reduction methods // Briefings in Bioinformatics. — 2016. — Vol. 17, no. 2. — P. 293—308.

метода, позволяющие учитывать популяционную стратификацию индивидуумов[26]. М. Парк и коллеги рассматривали MDR метод для многомерного фенотипа[27].

В данной диссертации мы продолжаем изучать и развивать метод MDR-EFE (Multifactor Dimensionality Reduction with Error Function Estimation). Впервые алгоритм был предложен[28] в работе А.В. Булинского и соавторов для исследования бинарного отклика, и получил дальнейшее развитие в последующих работах[29,30] [1-3,5,6]. Метод основан на статистической оценке функционала ошибки вида $Err(f) = |Y - f(X)|\psi(Y)$, где Y – изучаемый случайный отклик, X – вектор факторов, $f(\cdot)$ – предсказательная функция, а $\psi(\cdot)$ – штрафная функция. Оценка функционала ошибки строится по набору независимых одинаково распределенных векторов с помощью кросс-валидации для большей устойчивости алгоритма.

Цель работы. Целью работы является разработка новых методов идентификации значимых факторов с помощью функционала ошибки. В частности, ставится задача по построению модификаций предложенного ранее MDR-EFE метода на случай небинарной функции отклика. Также рассматривается модель с объясняющими факторами, имеющими абсолютно непрерывное распределение относительно меры Лебега в пространстве $\mathbb{R}^n$. Предлагается вариант MDR-EFE метода с последовательным отбором значимых переменных. Развивается теория перестановочных случайных величин. Одной из основных целей работы является изучение асимптотических свойств используемых оценок в предложенных модификациях MDR-EFE метода. Проводится компьютерное моделирование для иллюстрации работы MDR-EFE метода.

Структура и объем работы. Диссертация, объемом 110 страниц, состоит из введения, трех глав, заключения и списка литературы, насчитывающего 100 наименований. В заключении к диссертации сформулированы возможные направления дальнейшей деятельности.

В первой главе дается описание MDR-EFE метода идентификации значимых факторов с помощью функционала ошибки. Затем предлагается

[26] Abegaz F., Van Lishout F., Mahachie John J. M., Chiachoompu K., [et al.]. Performance of model-based multifactor dimensionality reduction methods for epistasis detection by controlling population structure // BioData Mining. — 2021. — Vol. 14, no. 1. — P. 1—20.

[27] Park M., Jeong H.-B., Lee J.-H., Park T. Spatial rank-based multifactor dimensionality reduction to detect gene–gene interactions for multivariate phenotypes // BMC bioinformatics. — 2021. — Vol. 22, no. 1. — P. 1—21.

[28] Bulinski A., Butkovsky O., Sadovnichy V., Shashkin A., [et al.]. Statistical Methods of SNP Data Analysis and Applications // Open Journal of Statistics. — 2012. — Vol. 2, no. 1. — P. 73—87.

[29] Bulinski A. Central limit theorem related to MDR-method // Asymptotic Laws and Methods in Stochastics. — 2015. — P. 113—128.

[30] Bulinski A., Kozhevin A. New version of the MDR method for stratified samples // Statistics, Optimization & Information Computing. — 2017. — Vol. 5, no. 1. — P. 1—18.

его модификация на случай небинарной функции отклика. Устанавливается критерий сильной состоятельности введенных оценок функционала ошибки в случае небинарной функции отклика. Доказывается теорема, которая обосновывает стратегию выбора набора значимых факторов. Доказывается теорема о сильной состоятельности функционала ошибки в случае объясняющих факторов, имеющих абсолютно непрерывное распределение относительно меры Лебега в пространстве $\mathbb{R}^n$.

Вторая глава посвящена изучению асимптотические свойства оценок функционала ошибки, построенных с помощью процедуры кросс-валидации. Доказывается центральная предельная теорема (ЦПТ) для регуляризованных оценок функционала ошибки в случае небинарной функции отклика. С целью получения дальнейших асимптотических результатов развивается теория перестановочных случайных величин. Доказывается аналог теоремы Эрдеша и Каца для перестановочных случайных величин. Устанавливается новый вариант ЦПТ для перестановочных случайных величин, с помощью которого доказывается новый вариант ЦПТ для оценок функционала ошибок. Полученные результаты о скорости сходимости построенных оценок к предельному распределению используются с целью получения асимптотических доверительных интервалов.

В третьей главе разрабатывается новая версия MDR-EFE метода с последовательным отбором значимых переменных. Для модели наивного байесовского классификатора устанавливаются оценки снизу для вероятности выбора значимого набора факторов MDR-EFE методом с последовательным отбором переменных. MDR-EFE реализуется в виде программного кода, а его работа иллюстрируется на данных компьютерного моделирования.

Научная новизна работы. Все результаты, представленные в диссертации, являются новыми.

Положения, выносимые на защиту:

1. Критерий сильной состоятельности оценки функционала ошибки в MDR-EFE методе для случая небинарной функции отклика.
2. Теорема, обосновывающая стратегию выбора набора значимых факторов.
3. Достаточные условия сильной состоятельности оценок в случае объясняющих факторов, имеющих абсолютно-непрерывное распределение относительно меры Лебега в пространстве $\mathbb{R}^n$.
4. ЦПТ для регуляризованных оценок функционала ошибки в случае небинарной функции отклика.
5. Новый вариант ЦПТ для серий перестановочных случайных величин. Новый вариант ЦПТ для оценок функционала ошибок.
6. Аналог теоремы Эрдеша и Каца для перестановочных случайных величин.

7. Оценки снизу для вероятности выбора значимого набора факторов MDR-EFE методом с последовательным отбором переменных в случае модели наивного байесовского классификатора.

Методы исследования. В работе используются классические методы теории вероятностей, вероятностные неравенства, асимптотические результаты для массивов случайных величин, анализ распределений случайных векторов. При доказательстве ЦПТ применяется техника перестановочных случайных величин. Часть теорем доказана с помощью результатов, справедливых для мартингалов.

Практическая и теоретическая значимость работы. Результаты диссертации носят теоретический характер. При этом они допускают и приложения. Разрабатываемый MDR-EFE метод и его модификации могут быть применимы в биостатистических задачах, требующих выявления факторов, оказывающих влияние на изучаемый отклик.

Апробация диссертации. Результаты диссертации докладывались на следующих **конференциях**.

1. International workshop «Probability, Analysis and Geometry», Ульм, Германия, 2013.
2. Международная научная конференция «Современные проблемы математики и механики», посвященная 75-летию академика РАН В.А. Садовничего, Москва, Россия, 2014.
3. XXI Международная научная конференция студентов, аспирантов и молодых ученых «Ломоносов», Москва, Россия, 2014.
4. XXXII International Seminar on Stability Problems for Stochastic Models, Трондхейм, Норвегия, 2014.
5. International Conference on Bioinformatics Models, Methods and Algorithms, Лиссабон, Португалия, 2015.
6. 6th Annual Canadian Human and Statistical Genetics Meeting, Квебек, Канада, 2017.
7. XXIV Международная научная конференция студентов, аспирантов и молодых ученых «Ломоносов», Москва, Россия, 2017.
8. V Международная конференция «Постгеном-2018», Казань, Россия, 2018.
9. The 5th International Conference on Stochastic Methods, Москва, Россия, 2020.
10. International conference «Limit Theorems of Probability Theory and Mathematical Statistics», Ташкент, Узбекистан, 2022.

Результаты диссертации неоднократно докладывались автором на следующих **научно-исследовательских семинарах**.

1. Большой семинар кафедры теории вероятностей под руководством академика РАН, профессора А.Н. Ширяева, механико-математический факультет, Московский государственный университет им. М.В.Ломоносова.

2. «Асимптотический анализ случайных процессов и полей» под руководством доктора физико-математических наук, профессора А.В. Булинского, механико-математический факультет, Московский государственный университет им. М.В.Ломоносова.
3. Аспирантский коллоквиум по теории вероятностей, математической статистике, теории случайных процессов под руководством академика РАН, профессора А.Н. Ширяева, механико-математический факультет, Московский государственный университет им. М.В.Ломоносва.
4. «Forschungsseminar Stochastische Geometrie und raumliche Statistik» под руководством Prof. E.Spodarev (Institut fur Stochastik, Ulm University, Germany, 2014 г.).

Публикации. Основные результаты диссертации изложены в 10 публикациях автора. Из них 4 статьи опубликованы в рецензируемых научных журналах, входящих в базы SCOPUS, Web of Science, RSCI. 2 статьи без соавторов опубликованы в трудах научных конференций. В материалах международных конференций представлены 4 публикации.

Личный вклад автора. Диссертантом совместно с научным руководителем проводился выбор темы, а также осуществлялось планирование всей работы. Профессору А.В. Булинскому принадлежит постановка задач и общий подход к их решению, им также доказаны леммы 2, 5, теоремы 8, 13 и следствия 1, 4. Предложение 2 и следствие 3 доказаны П. Алонсо-Руиз. Автору диссертации принадлежит доказательство остальных лемм, предложений, теорем, следствий, проведение компьютерного моделирования. В начале каждой главы диссертации также приводится список соответствующих публикаций с долей участия авторов.

Благодарность. В заключение автор выражает признательность научному руководителю профессору А.В. Булинскому за большую помощь в работе.

Содержание работы

Во введении к диссертации обсуждается практическая важность и актуальность задачи выявления значимых факторов, отмечается научная новизна работы, формулируются основные цели и задачи.

В первой главе рассматриваются обобщения MDR-EFE метода на случай небинарной функции отклика, а также на случай объясняющих факторов, имеющих абсолютно непрерывное распределение относительно меры Лебега в пространстве $\mathbb{R}^n$. В разделе 1.1 вводятся основные обозначения, а также определяются функционал ошибки Err, предсказательный алгоритм f_{PA} и их статистические оценки в случае бинарной функции отклика Y. Раздел 1.2 посвящен обобщению приведенных оценок для небинарной функции отклика Y. Кроме того, в этом разделе устанавливается

ряд результатов, включающих в себя критерий сильной состоятельности построенных оценок, теорему о применимости предлагаемого подхода к выявлению значимых наборов факторов и другие.

Пусть все случайные величины заданы на некотором вероятностном пространстве $(\Omega, \mathcal{F}, \mathbb{P})$. Интеграл Лебега от случайной величины $\xi : \Omega \to \mathbb{R}$ по мере $\mathbb{P}$ будем обозначать $\mathbb{E}(\xi)$. Пусть $X = (X_1, \ldots, X_n)$ – случайный вектор с компонентами $X_k : \Omega \to \{0,1,\ldots,s\}$, где $k = 1, \ldots, n$ и $s,n \in \mathbb{N}$. Положим $\mathbb{X} = \{0,\ldots,s\}^n$, $\mathbb{Y} = \{-m,\ldots,0,\ldots,m\}$, здесь $m \in \mathbb{N}$. Мы предполагаем, что $Y : \Omega \to \mathbb{Y}$, $f : \mathbb{X} \to \mathbb{Y}$ и штрафная функция $\psi : \mathbb{Y} \to \mathbb{R}_+$. Тривиальный случай $\psi \equiv 0$ исключается из рассмотрения.

Качество предсказаний Y значениями $f(X)$ исследуется с помощью функционала ошибки:

$$Err(f) := \mathbb{E}|Y - f(X)|\psi(Y).$$

Рассмотрим множества $A_y = \{x \in \mathbb{X} : f(x) = y\}$, где $y \in \mathbb{Y}$. Тогда мы можем представить $Err(f)$ в виде

$$Err(f) = \sum_{y,z \in \mathbb{Y}} |y - z|\psi(y)\mathbb{P}(Y = y, f(X) = z) = \sum_{z \in \mathbb{Y}} \sum_{x \in A_z} w^\top(x) q(z).$$

Здесь $q(z)$ – столбец с номером z матрицы Q размерности $(2m+1) \times (2m+1)$ с элементами $q_{y,z} = |y - z|$, $y,z \in \mathbb{Y}$ (элемент $q_{-m,-m}$ находится в левом верхнем углу Q),

$$w(x) = (\psi(-m)\mathbb{P}(Y = -m, X = x), \ldots, \psi(m)\mathbb{P}(Y = m, X = x))^\top,$$

и $\top$ обозначает транспонирование. Все векторы рассматриваются как столбцы. Функционал ошибки Err также можно представить в виде

$$Err(f) = \sum_{i=0}^{2m-1} \sum_{i-m<|y|\le m} \psi(y)\mathbb{P}(Y = y, |f(X) - y| > i).$$

Пусть $\xi^1, \xi^2, \ldots$ – последовательность независимых одинаково распределенных (н.о.р.) случайных векторов, имеющих такое же распределение, что и (X,Y). Для $N \in \mathbb{N}$ положим $\xi_N = (\xi^1, \ldots, \xi^N)$. Мы будем приближать $Err(f)$ используя ξ_N (при $N \to \infty$) и *предсказательный алгоритм* (PA). Данный предсказательный алгоритм использует функцию $f_{PA} = f_{PA}(x, \xi_N)$, заданную для $x \in \mathbb{X}$ и ξ_N и принимающую значения в $\mathbb{Y}$. Для $S \subset \{1, \ldots, N\}$ положим $\xi_N(S) = \{\xi^j, j \in S\}$ и $\overline{S} := \{1, \ldots, N\} \setminus S$. Для $K \in \mathbb{N}$, $(K > 1)$ введем разбиение множества $\{1, \ldots, N\}$ на подмножества

$$S_k(N) = \{(k-1)[N/K] + 1, \ldots, k[N/K]\mathbb{I}\{k < K\} + N\mathbb{I}\{k = K\}\},$$

где $k = 1, \ldots, K$, $[a]$ – целая часть числа $a \in \mathbb{R}$, $\mathbb{I}\{A\}$ – индикатор множества A. Пусть, как обычно, $\#A$ обозначает мощность конечного множества A.

Построим оценку введенного функционала ошибки $Err(f)$, основываясь на выборке ξ_N, алгоритме предсказания с f_{PA} и применяя K-кратную кросс-валидацию, где $K \in \mathbb{N}$, $K > 1$.

$$\widehat{Err}_K(f_{PA},\xi_N) := \sum_{i=0}^{2m-1} \sum_{i-m<|y|\le m} \frac{1}{K} \sum_{k=1}^{K} \sum_{j\in S_k(N)} \widehat{\psi}(y,\xi_N(S_k(N))) \times \frac{\mathbb{I}\{Y^j = y, |f_{PA}(X^j,\xi_N(\overline{S_k(N)})) - y| > i\}}{\sharp S_k(N)}. \quad (1)$$

Здесь для каждого $k \in \{1,\dots,K\}$ оценка $\widehat{\psi}(y,\xi_N(S_k(N))$ является сильно состоятельной оценкой $\psi(y)$ (при $N \to \infty$) для всех $y \in \mathbb{Y}$, т.е.

$$\widehat{\psi}(y,\xi_N(S_k(N))) \to \psi(y) \text{ п.н., } y \in \mathbb{Y}, \ N \to \infty. \quad (2)$$

В практических приложениях функция $\psi(\cdot)$ и ее статистическая оценка $\widehat{\psi}(\cdot,\cdot)$ выбираются таким образом, что соотношение (2) выполняется.

Мы хотим, чтобы сходимость (в определенном смысле) $f_{PA}(\cdot,\xi_N)$ к $f(\cdot)$ при $N \to \infty$ влекла соотношение

$$\widehat{Err}_K(f_{PA},\xi_N) \to Err(f) \quad \text{п.н., } \ N \to \infty. \quad (3)$$

Теорема 1. *Пусть $\xi^1,\xi^2,\dots$ – последовательность независимых одинаково распределенных случайных величин с таким же законом распределения, что и (X,Y), ψ – штрафная функция, $f : \mathbb{X} \to \mathbb{Y}$ и f_{PA} задают предсказательный алгоритм. Предположим, что существует такое непустое множество $U \subset \mathbb{X}$, что для каждого $x \in U$ и всех $k = 1,\dots,K$ имеем*

$$f_{PA}(x,\xi_N(\overline{S_k(N)})) \to f(x) \quad \text{п.н., } \ N \to \infty. \quad (4)$$

Тогда (3) *выполняется в том и только том случае, если*

$$\sum_{k=1}^{K} \sum_{x\in\mathbb{X}\setminus U} w^\top(x)\, Q\, \delta(N,x,k) \to 0 \ \text{ п.н., } \ N \to \infty, \quad (5)$$

где для $x \in \mathbb{X}$, $N \in \mathbb{N}$ и $k = 1,\dots,K$ вектор $\delta(N,x,k)$ имеет компоненты

$$\delta_y(N,x,k) = \mathbb{I}\{f_{PA}(x,\xi_N(\overline{S_k(N)})) = y\} - \mathbb{I}\{f(x) = y\}, \ y \in \mathbb{Y}.$$

Далее из теоремы 1 выводятся два следствия, относящиеся к условию (5).

Следствие 1. *Условие* (5) *Теоремы 1 эквивалентно следующему:*

$$\sum_{k=1}^{K} \sum_{t\in\mathbb{Y}} \sum_{x\in\mathbb{X}(t,U)} L^\top(x) I(N,x,k,t) \to 0 \ \text{ п.н., } \ N \to \infty, \quad (6)$$

где $L^\top(x) := (L_{-m+1}(x),\ldots,L_m(x))$ *– вектор с компонентами* $L_y(x) := w_{-m}(x)+\ldots+w_{y-1}(x)-w_y(x)-\ldots-w_m(x), \quad y=-(m-1),\ldots,m,$ *и* $\mathbb{X}(t,U) := (\mathbb{X}\setminus U)\cap\{x\in M : f(x)=t\}$.

Следствие 2. *Пусть* ψ *– штрафная функция,* $f:\mathbb{X}\to\mathbb{Y}$, *и предсказательный алгоритм определяется функцией* f_{PA}. *Предположим, что для некоторого множества* $U\subset\mathbb{X}$ *выполняется условие* (4). *Будем считать, что для каждого* $t\in\mathbb{Y}$ *и произвольного* $x\in\mathbb{X}(t,U)$ *существуют* $i=i(x)$, $j=j(x)$, *принадлежащие* $\mathbb{Y}$, $i<j$, *такие, что*

$$i\le f_{PA}(x,\xi_N(\overline{S_k(N)}))\le j \;\; \text{п.н. для } k=1,\ldots,K$$

при достаточно больших N. *Тогда условие*

$$L_{\min\{t,i\}+1}(x)=\ldots=L_{\max\{t,j\}}(x)=0$$

влечет выполнение (6).

Далее приводятся примеры оценок $\widehat{\psi}(y,\xi_N(S_k(N))$ и $f_{PA}(x,\xi_N(\overline{S_k(N)}))$, для которых условия теоремы 1 выполняются.

Для многих моделей естественным предположением является то, что функция отклика Y зависит лишь от некоторых факторов $X_{k_1},\ldots,X_{k_r}$, где $1\le k_1<\ldots<k_r\le n$. Другими словами, для произвольных $x=(x_1,\ldots,x_n)\in M$ и $y\in\mathbb{Y}$,

$$\mathbb{P}(Y=y|X_1=x_1,\ldots,X_n=x_n)=\mathbb{P}(Y=y|X_{k_1}=x_{k_1},\ldots,X_{k_r}=x_{k_r}),$$

если $P(X_1=x_1,\ldots,X_n=x_n)\ne 0$. Любой такой набор индексов $\{k_1,\ldots,k_r\}$ будем называть *значимым*. Для произвольного набора факторов β водятся оценки предсказательного алгоритма $f_{PA}=\widehat{f}^{\beta}_{PA}$, использующие лишь те компоненты вектора X, индексы которых содержатся в β.

Теорема 3. *Пусть* $\alpha=(k_1,\ldots,k_r)$ *является значимым набором* $\{k_1,\ldots,k_r\}\subset\{1,\ldots,n\}$. *Тогда для произвольного* $\varepsilon>0$ *и всех* $\beta=(m_1,\ldots,m_r)$ *с* $\{m_1,\ldots,m_r\}\subset\{1,\ldots,n\}$ *выполняется следующее неравенство:*

$$\widehat{Err}_K(\widehat{f}^{\alpha}_{PA})\le\widehat{Err}_K(\widehat{f}^{\beta}_{PA})+\varepsilon \;\; \text{п.н.} \tag{7}$$

для всех достаточно больших N.

Из теоремы 3 вытекает, что для дальнейшего анализа естественно в качестве значимого набора выбирать такие наборы $\{k_1,\ldots,k_r\}\subset\{1,\ldots,n\}$, что $\widehat{Err}_K(\widehat{f}^{\alpha}_{PA})$ с $\alpha=(k_1,\ldots,k_r)$ имеет минимальное (или почти минимальное) значение среди всех $\widehat{Err}_K(\widehat{f}^{\beta}_{PA})$, где $\beta=(m_1,\ldots,m_r)$ и $\{m_1,\ldots,m_r)\subset\{1,\ldots,n\}$. Важно отметить, что мы установили сходимость почти наверное в (7), так как нам необходимо сравнивать $\widehat{Err}_K(\widehat{f}^{\beta}_{PA})$ для различных $\beta=(m_1,\ldots,m_r)$ одновременно. Если бы была доказана только

сходимость по вероятности, то потребовалось бы использовать неравенство Бонферрони. Это не позволило бы осуществить с большой вероятностью одновременное сравнение многих наборов факторов. Также заметим, что в целом ряде задач имеется набор из небольшого количества объясняющих факторов $X_{k_1},\ldots,X_{k_r}$, в то время как общее количество значимых факторов $X_1,\ldots,X_n$ может быть достаточно велико.

В разделе 1.3 рассматривается случай объясняющих факторов, имеющих абсолютно-непрерывное распределение. Пусть теперь X принимает значения в $\mathbb{X}=\mathbb{R}^n$. Будем считать, что у случайного вектора X есть плотность ρ по мере Лебега в $\mathbb{R}^n$. Положим $M=\{x\in\mathbb{X}:\rho_X(x)>0\}$. В данном разделе рассматривается бинарная функция отклика Y с $\mathbb{Y}=\{-1,1\}$, в случае которой предсказательный алгоритм строится следующим образом:

$$f_{PA}(x,\xi_N)=\begin{cases}1, & \widehat{\mathbb{P}}(Y=1|X=x)>\widehat{\gamma}(\psi),\\ -1, & \text{иначе},\end{cases} \tag{8}$$

где $\widehat{\mathbb{P}}(Y=1|X=x)$ – оценка условной вероятности $\mathbb{P}(Y=1|X=x)$, а $\widehat{\gamma}(\psi)$ – оценка пороговой функции $\gamma(\psi):=\psi(-1)/(\psi(-1)+\psi(1))$.

Теорема 4. *Пусть $\xi^1,\xi^2,\ldots$ – последовательность н.о.р. случайных векторов, имеющих тот же закон распределения, что и вектор (X,Y), ψ – некоторая штрафная функция, $f:\mathbb{X}\to\{-1,1\}$ и f_{PA} задает алгоритм предсказания согласно (8). Предположим, что при $N\to\infty$ последовательность случайных величин $\widehat{\gamma}_N(\psi)$ сходится вполне к $\gamma(\psi)$. Пусть у вектора X существует плотность $\rho(x)$, $x\in\mathbb{R}^n$. Кроме того, пусть функция*

$$\kappa(\delta)=\int_{\mathbb{R}^n}\mathbb{I}\left\{x:\mathbb{P}(Y=1|X=x)\in[\gamma(\psi)-\delta,\gamma(\psi)+\delta]\right\}\rho(x)dx$$

непрерывна в окрестности точки $\delta=0$. Потребуем, чтобы $\kappa(0)=0$. Кроме того, пусть последовательность случайных величин $\sup_{x\in\mathbb{R}^n}|\widehat{\mathbb{P}}_N(Y=1|X=x)-\mathbb{P}(Y=1|X=x)|$ сходилась к 0 вполне при $N\to\infty$, то есть, выполняется неравенство

$$\sum_{N=1}^{\infty}\mathbb{P}\left(\sup_{x\in\mathbb{R}^n}|\widehat{\mathbb{P}}_N(Y=1|X=x)-\mathbb{P}(Y=1|X=x)|>\delta\right)<\infty$$

для произвольного $\delta>0$. Тогда справедливо соотношение

$$\widehat{Err}_K(f_{PA},\xi_N)\to Err(f)\quad \text{п.н., } N\to\infty.$$

В работе[31] А.В. Булинского в замечании 4 объясняется, почему предложенная в работе[32] Д. Велеса штрафная функция

$$\psi(y) = \frac{c}{\mathbb{P}(Y = y)}, \quad y \in \{-1,1\}, \quad c > 0, \tag{9}$$

является хорошим выбором. Если штрафная функция задана согласно (9), то справедливо

$$\gamma(\psi) = \frac{\psi(-1)}{\psi{-}1 + \psi(1)} = \frac{c/\mathbb{P}(Y = -1)}{c/\mathbb{P}(Y = -1) + c/\mathbb{P}(Y = 1)} = \mathbb{P}(Y = 1).$$

В качестве оценки вероятности $\mathbb{P}(Y = 1)$ естественно использовать

$$\widehat{\gamma}_N(\psi,\xi_N(\overline{S_k(N)})) := \frac{1}{\#\overline{S_k(N)}} \sum_{j \in \overline{S_k(N)}} \mathbb{I}\left\{Y^j = 1\right\}, \quad k = 1,\ldots,K. \tag{10}$$

Предложение 1. *Предположим, что штрафная функция $\psi(y)$ задана согласно* (9)*, а оценка $\widehat{\gamma}_N(\psi, \xi_N(\overline{S_k(N)}))$, $k = 1,\ldots,K$, определена в* (10)*. Тогда для любого положительного δ справедливо неравенство*

$$\mathbb{P}(|\widehat{\gamma}_N(\psi, \xi_N(\overline{S_k(N)})) - \gamma(\psi)| \geq \delta) \leq 2\exp(-2\delta^2 \#\overline{S_k(N)}), \quad k = 1,\ldots,K.$$

В частности, из этого следует, что для любого $k = 1,\ldots,K$ оценка $\widehat{\gamma}_N(\psi, \xi_N(\overline{S_k(N)}))$ сходится к $\gamma(\psi)$ вполне.

Следующая теорема опирается на результаты статьи[33] Л. Девроя. В указанной работе исследуются оценки плотности $\rho(\cdot)$ (по мере Лебега) вектора со значениями в $\mathbb{R}^n$, построенные по выборке независимых одинаково распределенных векторов $X_1,\ldots,X_N$ с плотностью ρ на $\mathbb{R}^n$. Авторы рассматривают ядерную оценку специального вида:

$$\widehat{\rho}(x,k,N) := \frac{1}{N}\sum_{i=1}^{N} \frac{1}{(H_{N,k,i})^n} K\left(\frac{X_i - x}{H_{N,k,i}}\right), \; x \in \mathbb{R}^n, \; N \in \mathbb{N}, \tag{11}$$

где $N > 1$, $H_{N,k,i}$ – расстояние от X_i до $k-$го ближайшего соседа среди X_j, $j \neq i$, $k = k_N$ – положительные целые числа. Далее мы будем опускать индекс k в обозначениях $\rho(x,k,N)$ и $H_{N,k,i}$. В работе Л. Девроя и соавторов доказывается следующий важный результат.

[31] Bulinski A. On foundation of the dimensionality reduction method for explanatory variables // Journal of Mathematical Sciences. — 2014. — Vol. 199, no. 2. — P. 113—122.

[32] Velez D. R., White B. C., Motsinger A. A., Bush W. S., [et al.]. A balanced accuracy function for epistasis modeling in imbalanced datasets using multi-factor dimensionality reduction // Genetic Epidemiology. — 2007. — Vol. 31, no. 4. — P. 306—315.

[33] Devroye L., Penrod C. S. The strong uniform convergence of multivariate vari able kernel estimates // Canadian Journal of Statistics. — 1986. — Vol. 14, no. 3. — P. 211—220.

Теорема 5 (Devroye L., Penrod C.S., 1986). *Если*

$$K(x) = \mathbb{I}\{||x|| \le 1/c\}, \tag{12}$$

где c – положительная константа,

$$\lim_{N\to\infty} \frac{k}{N} = 0, \tag{13}$$

$$\lim_{N\to\infty} \frac{k}{\log N} = \infty, \tag{14}$$

и функция ρ равномерно непрерывна, то оценка плотности, заданная в (11) *с помощью ядра* (12)*, обладает следующим свойством. Для всех $\varepsilon > 0$ существуют $\delta > 0$ и N_0 такие, что*

$$\mathbb{P}(\sup_x |\widehat{\rho}(x,N) - \rho(x)| > \varepsilon) \le \exp(-\delta k), \qquad N > N_0.$$

Опираясь на теорему **5**, строятся оценки для $\mathbb{P}(Y = 1|X = x)$:

$$\widehat{\mathbb{P}}(Y = 1|X = x) = \frac{\widehat{f}_{Y|X}(x|1,N)\widehat{\mathbb{P}}_N(Y = 1)}{\widehat{\rho}_X(x,N)}.$$

Доказывается следующее утверждение.

Теорема 6. *Пусть вектор X имеет абсолютно-непрерывную плотность $\rho_X(x)$ с носителем M. Пусть функция κ из* (**4**) *непрерывна в окрестности 0 и $\kappa(0) = 0$. Предположим, что для некоторых положительных констант C_1,C_2*

$$C_1 < \rho_X(x) < C_2, \quad \forall x \in M.$$

$$C_1 < f_{X|Y}(x|1) < C_2, \quad \forall x \in M.$$

Кроме того, будем считать, что выполнены условия (12) - (14)*. Тогда справедливо соотношение*

$$\widehat{Err}_K(f_{PA},\xi_N) \to Err(f) \quad \text{п.н., } N \to \infty.$$

Во второй главе исследуются асимптотические свойства оценок функционала ошибки с использованием процедуры кросс-валидации. В разделе 2.1 вводятся *регуляризованные версии* оценок предсказательного алгоритма $\widehat{f}^{\beta}_{PA,\varepsilon}$. Доказывается следующая теорема.

Теорема 7. *Пусть $\varepsilon_N \to 0$ и $N^{1/2}\varepsilon_N \to \infty$ при $N \to \infty$. Тогда для каждого $K \in \mathbb{N}$, произвольного вектора $\beta = (m_1, \dots m_r)$ с $1 \le m_1 < \dots < m_r \le n$, соответствующей функции $f = f^{\beta}$ и предсказательного алгоритма $f_{PA} = \widehat{f}^{\beta}_{PA,\varepsilon}$ выполняется следующее соотношение:*

$$\sqrt{N}(\widehat{Err}_K(f_{PA},\xi_N) - Err(f)) \xrightarrow{law} Z \sim N(0,\sigma^2), \ N \to \infty.$$

Здесь σ^2 – дисперсия случайной величины

$$V=\sum_{i=0}^{2m-1}\sum_{i-m<|y|\le m}\frac{\mathbb{I}\{Y=y\}}{\mathbb{P}(Y=y)}\left(\mathbb{I}\{|f(X)-y|>i\}-\mathbb{P}(|f(X)-y|>i\big|Y=y)\right). \tag{15}$$

В разделе 2.2 вводится понятие перестановочных случайных величин. Пусть $\Pi(n)$ обозначает множество перестановок элементов $\{1,\dots,n\}$. Последовательность случайных величин $\{X_n\}_{n\in\mathbb{N}}$ называется перестановочной, если для любого $n\in\mathbb{N}$, $X_1,\dots,X_n$ перестановочны, то есть для каждой перестановки $\pi\in\Pi(n)$,

$$Law(X_1,\dots,X_n)=Law(X_{\pi(1)},\dots,X_{\pi(n)}).$$

Иначе говоря, $(X_1,\dots,X_n)\stackrel{d}{=}(X_{\pi(1)},\dots,X_{\pi(n)})$. Устанавливается аналог теоремы Эрдеша и Каца для последовательности перестановочных случайных величин. Определим $G_\mu\colon\mathbb{R}\to\mathbb{R}$ формулой

$$G_\mu(x):=\int_{\mathfrak{F}}\mathbb{I}_{(0,\infty)}(\sigma_F^2)G(x/\sigma_F)\,\mu(dF), \tag{16}$$

где $G(x):=(2\Phi(x)-1)\mathbb{I}_{[0,\infty)}(x),\ x\in\mathbb{R}$, а μ – вероятностное распределение управляющей случайной меры из теоремы де Финетти[34]. Устанавливается следующий результат.

Теорема 10. *Пусть $\{X_n\}_{n\in\mathbb{N}}$ – такая последовательность перестановочных случайных величин с нулевым средним и дисперсией $0<\mathbb{E}X_1^2<\infty$, что $\mathbb{E}X_1X_2=0$. Тогда*

$$\lim_{n\to\infty}\mathbb{P}(\max(S_1,\dots,S_n)<x\sqrt{n})=\mathbb{P}(\sigma_F^2=0)\mathbb{I}_{[0,\infty)}(x)+G_\mu(x),$$

где $G_\mu\colon\mathbb{R}\to\mathbb{R}$ введено в (16), μ – распределение управляющей случайной меры последовательности $\{X_n\}_{n\in\mathbb{N}}$.

В разделе 2.3 устанавливается новая версия центральной предельной теоремы для массива перестановочных случайных величин. Для массива $\{X_{n,i},1\le i\le k_n,n\in\mathbb{N}\}$ будем использовать следующие обозначения

$$\widehat{\mu}_{k_n}:=\frac{1}{k_n}\sum_{i=1}^{k_n}X_{n,i},\quad \widehat{\sigma}^2_{k_n}:=\frac{1}{k_n}\sum_{i=1}^{k_n}(X_{n,i}-\widehat{\mu}_{k_n})^2. \tag{17}$$

Доказываются следующие леммы.

[34]De Finetti B. Funzione caratteristica di un fenomeno aleatorio // Atti del Congresso Internazionale dei Matematici: Bologna del 3 al 10 de settembre di 1928. — 1929. — P. 179—190.

Лемма 5. *Пусть* $\{X_{n,i}, 1 \le i \le k_n, n \in \mathbb{N}\}$ – *построчно перестановочный массив случайных величин, где положительные числа* $k_n \to \infty$ *при* $n \to \infty$. *Предположим, что*

1°. $\sup_{n\in\mathbb{N}} \mathbb{E}X_{n,1}^4 < \infty$,

2°. $\mathbb{E}X_{n,1}^2 - \mathbb{E}X_{n,1}X_{n,2} \to \sigma^2 > 0$, $n \to \infty$,

3°. $\mathsf{cov}(X_{n,1}^2, X_{n,2}^2) + \mathsf{cov}(X_{n,1}X_{n,2}, X_{n,3}X_{n,4}) - 2\,\mathsf{cov}(X_{n,1}^2, X_{n,2}X_{n,3}) \to 0$,

при $n \to \infty$.

Тогда, для любой последовательности положительных чисел $(m_n)_{n\in\mathbb{N}}$ *такой, что* $m_n \to \infty$ *и* $m_n/k_n \to \alpha < 1$ *при* $n \to \infty$, *выполняется следующее соотношение:*

$$\frac{1}{\sqrt{m_n}} \sum_{i=1}^{m_n} (X_{n,i} - \widehat{\mu}_{k_n}) \xrightarrow{law} Z_{0,(1-\alpha)\sigma^2} \sim \mathcal{N}(0,(1-\alpha)\sigma^2), \ n \to \infty.$$

Пусть $K \in \mathbb{N}$, и предположим, что $N/K = q$, где $q \in \mathbb{N}$. Тогда $\sharp S_k(N) = q$ для каждого $k = 1,\ldots,K$. Рассмотрим последовательность $K \times q$-матриц $(C^{(N)})_{N\in\mathbb{N}}$ с элементами

$$\xi_{k,j}^{(N)} := \sum_{i=0}^{2m-1} \sum_{i-m<|y|\le m} \widehat{\psi}(y,S_k(N)) \cdot \mathbb{I}\{Y^{j+(k-1)q} = y, |f(X^{j+(k-1)q}) - y| > i\},$$

где $k = 1,\ldots,K$ и $j = 1,\ldots,q$. Введем

$$X_{N,j} := \frac{1}{\sqrt{K}} \sum_{k=1}^{K} \xi_{k,j}^{(N)}, \quad j = 1,\ldots,q. \tag{18}$$

Рассмотрим треугольный массив $\{X_{N,i}, 1 \le i \le q, N \in \mathbb{N}\}$ с элементами, определенными в (18). Положим $k_n = q$ в лемме 1 и будем писать N вместо n.

Лемма 6. *Предположим, что для каждого* $N \in \mathbb{N}$, *произвольного* $y \in \mathbb{Y}$ *и всех* $k = 1,\ldots,K$

$$\sup_{y\in\mathbb{Y},\ N\in\mathbb{N},\ k\in\{1,\ldots,K\}} \mathbb{E}\left(\widehat{\psi}(y,S_k(N))\right)^4 < \infty. \tag{19}$$

Пусть $(m_N)_{N\in\mathbb{N}}$ – *последовательность положительных целых чисел таких, что* $m_N \le q$, $m_N \to \infty$ *и* $m_N/N \to \alpha < 1$ *при* $N \to \infty$. *Тогда*

$$\frac{1}{\sqrt{m_N}} \sum_{i=1}^{m_N} (X_{N,i} - \widehat{\mu}_N) \xrightarrow{law} Z_{0,(1-\alpha)\sigma^2} \sim \mathcal{N}(0,(1-\alpha)\sigma^2),$$

где μ_N *введено в* (17) (*с* $k_n = q$ *и* N *вместо* n) *и*

$$\sigma^2 = \mathbb{E}\Big[\sum_{i=0}^{2m-1} \sum_{i-m<|y|\le m} \psi(y)(\mathbb{I}\{Y=y, |f(X)-y|>i\} - \mathbb{P}(Y=y, |f(X)-y|>i))\Big]^2. \quad (20)$$

Для последовательности случайных величин $(\eta_N)_{N\in\mathbb{N}}$ мы будем писать $\eta_N = O_{\mathbb{P}}(1)$, если для всех $\gamma > 0$ существует $M(\gamma) > 0$ такая, что $\mathbb{P}(|\eta_N| \ge M(\gamma)) \le \gamma$ для всех достаточно больших N. Пусть $(m_N)_{N\in\mathbb{N}}$ – последовательность положительных целых чисел таких, что $m_N \le q$ для $q = [N/K]$, и

$$m_N \to \infty,\ m_N/N \to 0,\ \text{когда}\ N \to \infty.$$

Теорема 12. *Пусть* $(m_N)_{N\in\mathbb{N}}$ – *последовательность, определенная выше. Предположим, что* $\varepsilon = (\varepsilon_N)_{N\in\mathbb{N}}$ – *такая последовательность положительных чисел, что* $\varepsilon_N \to 0$ *и* $m_N^{1/2}\varepsilon_N \to \infty$ *при* $N \to \infty$. *Рассмотрим произвольный вектор* $\beta = (k_1, \dots k_r)$ *с* $1 \le k_1 < \dots < k_r \le n$, *соответствующую функцию* $f = f^\beta$ *и предсказательный алгоритм* $f_{PA} = \widehat{f}^{\beta}_{PA,\varepsilon}$. *Пусть для всех* $y \in \mathbb{Y}$ *и* $k \in \{1, \dots, K\}$ *оценка* $\widehat{\psi}(y, S_k(N))$ *сильно состоятельна и*

$$\sqrt{\sharp S_k(N)}(\widehat{\psi}(y, S_k(N)) - \psi(y)) = O_{\mathbb{P}}(1),\ N \to \infty.$$

Пусть также выполнено (19). *Тогда при* $N \to \infty$ *справедливо соотношение*

$$\sqrt{m_N}\Big(\frac{1}{m_N}\sum_{j=1}^{m_N}\frac{1}{\sqrt{K}}\sum_{k=1}^{K}\sum_{i=0}^{2m-1}\sum_{i-m<|y|\le m}\Big[\widehat{\psi}(y, S_k(N))\mathbb{I}\{A_N(i,j,k,y)\}\Big] - Err(f)\Big) \xrightarrow{law} Z_{0,\sigma^2}.$$

Здесь $A_N(i,j,k,y) = \{Y^{j+(k-1)q} = y, |f_{PA}(X^{j+(k-1)q}) - y| > i\}$, $Z_{0,\sigma^2} \sim \mathcal{N}(0,\sigma^2)$ *и* σ^2 *определена в* (20).

С целью упрощения обозначений мы будем писать в следующей теореме $\widehat{Err}_K(f_{PA}, \xi_N)$ для случайной величины, введенной в (1), заменяя $\widehat{\psi}(y, S_k(N))$ на $\widehat{\psi}(y, \overline{S_k(N)})$, $y \in \mathbb{Y}$, $k = 1, \dots, K$.

Теорема 13. *Пусть* $\varepsilon_N \to 0$ *и* $N^{1/2}\varepsilon_N \to \infty$ *при* $N \to \infty$. *Если для любого* $K \in \mathbb{N}$ *и произвольного вектора* $\beta = (k_1, \dots k_r)$ *с* $1 \le k_1 < \dots < k_r \le n$ *соответствующие* $f = f^\beta$ *и предсказательный алгоритм определяются как* $f_{PA} = \widehat{f}^{\beta}_{PA,\varepsilon}$, *то выполняется следующее соотношение:*

$$\sqrt{N}(\widehat{Err}_K(f_{PA}, \xi_N) - Err(f)) \xrightarrow{law} Z_{0,\sigma^2} \sim \mathcal{N}(0,\sigma^2),\ N \to \infty. \quad (21)$$

Здесь σ^2 – *дисперсия случайной величины* V, *заданной в* (15).

В третьей главе рассматривается MDR-EFE метод с последовательным отбором переменных. Предполагается, что выполняются следующие *условия*:

1. $X_i : \Omega \to \{0,1\}$, $i = 1, \ldots, n$; $Y : \Omega \to \{-1,1\}$;
2. Y зависит только от $X_1, \ldots, X_r$ для некоторого натурального $r < n$ и не зависит от $X_{r+1}, \ldots, X_n$;
3. $\mathbb{P}(X|Y) = \prod_{i=1}^{n} \mathbb{P}(X_i|Y)$.

Для упрощения обозначений мы пишем $\mathbb{P}(X|Y)$ вместо $\mathbb{P}(X = x|Y = y)$, $x \in \mathbb{X}$, $y \in \mathbb{Y}$. Далее мы будем использовать обозначения

$$\theta_{v,y}^{(i)} := \mathbb{P}(X_i = v|Y = y),$$

где $v \in \{0,1\}$, $y \in \{-1,1\}$. Из введенных выше условий следует, что данные удовлетворяют модели логистической регрессии:

$$logit(\mathbb{P}(Y = 1|X)) = \beta_0 + \sum_{i=1}^{n} \beta_i X_i.$$

Будем считать, что нам известно число значимых факторов r. На первом шаге находится оценка функционала ошибки для всех наборов из одного фактора и выбирается фактор X_{i_1} с наименьшей ошибкой $\widehat{Err}_K(\widehat{f}_{PA}^{(i_1)})$. На следующем шаге рассматриваются наборы из двух факторов вида $(X_{i_1}, \cdot)$, и выбирается такой фактор X_{i_2}, что среди всех указанных пар набор (X_{i_1}, X_{i_2}) имеет наименьшую ошибку. Аналогично, на каждом из оставшихся $r - 2$ шагов добавляется по одному фактору. В итоге мы получаем набор $(X_{i_1}, \ldots, X_{i_r})$, который считаем значимым. Естественным образом возникает вопрос, какова вероятность того, что набор $(X_{i_1}, \ldots, X_{i_r})$ будет состоять из значимых факторов и только из них. Далее устанавливаются оценки для упомянутой вероятности.

Без ограничения общности будем считать, что $\beta_1 > \beta_2 > \ldots > \beta_r > 0$, а также $\beta_j = 0$ при $j = r + 1, \ldots, n$. Пусть s_i – номер фактора, выбранного в качестве значимого на i-ом шаге алгоритма ($i = 1, \ldots, r$). Для описанной выше модели наибольшее значение должна иметь вероятность $\mathbb{P}(s_1 = 1, \ldots, s_r = r)$ при достаточно больших N. Для данной вероятности мы получим оценку снизу и увидим, что она стремится к 1 при росте числа наблюдений N.

Теорема 14. *Пусть выполнены условия, достаточные для справедливости утверждения теоремы 7. Тогда*

$$\mathbb{P}(s_1 = 1, \ldots, s_r = r) \geq \prod_{k=1}^{r} \Big(1 - \frac{1}{N} \sum_{t=k+2}^{n} \frac{8V_{max} + o(1)}{\big(c_{k+1,t}^{(1,\ldots,k)} + o(1/\sqrt{N})\big)^2}\Big), N \to \infty, \tag{22}$$

где V_{max} и $c_{k+1,t}^{(1,\ldots,k)}$ – некоторые положительные константы.

Из доказанной теоремы легко выводится следующий результат.

Следствие 5. *Пусть выполнены условия теоремы* ***14***. *Пусть* C_1,C_2 – *некоторые константы. Тогда для достаточно больших* N *справедливы неравенства*

$$\left(1-\frac{C_1C_2}{N}\right)^r \le \mathbb{P}(s_1=1,\dots,s_r=r) \le \left(1-\frac{C_1}{N}\right)^r$$

Наконец, доказывается следующая теорема.

Теорема 15. *В* (22) *константы* $c_{k+1,t}^{(1,\dots,k)}$ *могут быть определенным образом выражены через коэффициенты логистической регрессии.*

В разделе 3.3 содержится описание реализации MDR-EFE метода в виде программного кода на языке R. Данные компьютерного моделирования используются для сравнения работы MDR-EFE метода, логистической регрессии и классического MDR метода.

Заключение. Тематика диссертации относится к области разработки вероятностно-статистических методов выявления значимых факторов, влияющих на изучаемую случайную функцию отклика. Проведено исследование асимптотических свойств MDR-EFE (multifactor dimensionality reduction – error function estimation) метода и его модификаций. Перечислим основные результаты диссертации.

В случае небинарной функции отклика построена оценка функционала ошибки в MDR-EFE методе по имеющейся выборке с помощью процедуры кросс-валидации. Установлен критерий сильной состоятельности построенной оценки. Доказана теорема, обосновывающая возможность использования оценок функционала ошибки для выбора наборов значимых факторов. Найдены достаточные условия сильной состоятельности оценок функционала ошибки в случае объясняющих факторов, имеющих абсолютно непрерывное распределение относительно меры Лебега в пространстве $\mathbb{R}^n$. Показано, что построенные ядерные оценки условных плотностей с помощью статистик k ближайших соседей удовлетворяют условиям полученной теоремы. Доказана центральная предельная теорема для регуляризованных версий оценок функционала ошибки в случае небинарной функции отклика. С помощью техники перестановочных случайных величин доказан новый вариант центральной теоремы для оценок функционала ошибки в MDR-EFE методе. Установлен аналог теоремы Эрдеша и Каца для перестановочных случайных величин.

Разработан вариант MDR-EFE метода с последовательным отбором переменных. В случае модели наивного байесовского классификатора получены оценки снизу для вероятности выбора значимого набора факторов MDR-EFE методом с последовательным отбором переменных. Применение MDR-EFE метода проиллюстрировано на данных компьютерного моделирования.

Дальнейшие исследования по тематике диссертации могут проводиться в направлении доказательства критерия сильной состоятельности оценки функционала ошибки в случае функции отклика, имеющей абсолютно-непрерывное распределение, оценок скорости сходимости в центральной предельной теореме для оценок функционала ошибок в MDR-EFE методе. Другой задачей может быть вывод оценок вероятности выбора значимого набора факторов для MDR-EFE метода с последовательным отбором переменных в различных моделях (отличных от модели наивного байесовского классификатора), а также исследование моделей, в которых количество значимых факторов неизвестно.

Публикации автора по теме диссертации

Статьи в научных журналах Web of Science, SCOPUS, RSCI

1. Bulinski A. V., Rakitko A. S. Estimation of nonbinary random response // Doklady Mathematics. — 2014. — Vol. 89, no. 2. — P. 225—229.

   ИФ WoS (JIF) - 0.486 / 0.31 п.л. / вклад соискателя 0.22 п.л.
   В публикации А.В. Булинскому принадлежит постановка задач и общий подход к их решению, им также доказана лемма 1, все остальные результаты доказаны автором диссертации.

2. Bulinski A., Rakitko A. MDR method for nonbinary response variable // Journal of Multivariate Analysis. — 2015. — Vol. 135. — P. 25—42.

   ИФ WoS (JIF) - 1.387 / 1.13 п.л. / вклад соискателя 0.91 п.л.
   В публикации А.В. Булинскому принадлежит постановка задач и общий подход к их решению, им также доказана лемма 2 (лемма 2 в диссертации), следствие 1 (следствие 1 в диссертации) и следствие 4 (теорема 8 в диссертации), все остальные результаты доказаны автором диссертации.

3. Bulinski A., Rakitko A. Simulation and analytical approach to the identification of significant factors // Communications in Statistics Part B: Simulation and Computation. — 2016. — Vol. 45, no. 5. — P. 1430—1450.

   ИФ WoS (JIF) - 1.162 / 1.31 п.л. / вклад соискателя 1.05 п.л.
   В публикации А.В. Булинскому принадлежит постановка задач и общий подход к их решению, им также доказана лемма 1 (лемма 5 в диссертации), теорема 3 (теорема 13 в диссертации) и следствие 1 (следствие 4 в диссертации), все остальные результаты доказаны автором диссертации.

4. Ruiz P. A., Rakitko A. The limit theorem for maximum of partial sums of exchangeable random variables // Statistics and Probability Letters. — 2016. — Vol. 119. — P. 357—362.

   ИФ WoS (JIF) - 0.718 / 0.38 п.л./ вклад соискателя 0.19 п.л.
   В публикации П. Алонсо-Руиз принадлежит предложение 1 (предложение 2 в диссертации) и следствие 1 (следствие 3 в диссертации), все остальные результаты доказаны автором диссертации.

Статьи в трудах научных конференций

5. Rakitko A. S. MDR-EFE method with forward selection // The 5th International Conference on Stochastic Methods (ICSM-5). — 2020. — P. 163—167.

6. Rakitko A. S. Multifactorial Dimensionality Reduction for Disordered Trait // Proceedings of the International Joint Conference on Biomedical Engineering Systems and Technologies. Vol. 3. — 2015. — P. 232—236.

Тезисы докладов в материалах научных конференций

7. Ракитько А. С. Последовательный отбор переменных в MDR-EFE методе // Сборник тезисов XXIV Международной научной конференции студентов, аспирантов и молодых учёных «Ломоносов–2017». — Макс Пресс Москва, 2017. — С. 1—2.

8. Ракитько А. С. Центральные предельные теоремы для массивов перестановочных случайных величин // Сборник тезисов XXI Международной конференции студентов, аспирантов и молодых учёных «Ломоносов-2014». — 2014. — С. 1—2.

9. Rakitko A. S. Multifactor dimensionality reduction method and simulation techniques // Abstracts of XXXII International Seminar on Stability Problems for Stochastic Models and VIII International Workshop «Applied Problems in Theory of Probabilities and Mathematical Statistics related to modeling of information systems». — Institute of Informatics Problems, Russian Academy of Sciences, 2014. — P. 94—95.

10. Rakitko A. S. On the application of MDR-EFE method for relevant feature selection // Abstract of communications for international conference «Limit Theorems of Probability Theory and Mathematical Statistics». — 2022. — P. 98.

Ракитъко Александр Сергеевич

Идентификация значимых факторов с помощью функционала ошибки

Автореф. дис. на соискание ученой степени канд. физ.-мат. наук

Подписано в печать ___.___._____. Заказ № _____
Формат 60×90/16. Усл. печ. л. 1. Тираж 100 экз.
Типография _____