Московский государственный университет
имени М.В.Ломоносова

На правах рукописи

Кожевин Алексей Александрович

Вероятностные методы отбора значимых факторов

01.01.05 - теория вероятностей и математическая статистика

АВТОРЕФЕРАТ
диссертации на соискание ученой степени
кандидата физико-математических наук

Москва, 2021 г.

Работа выполнена на кафедре теории вероятностей механико-математического факультета МГУ имени М.В.Ломоносова.

Научный руководитель: **Булинский Александр Вадимович**
доктор физико-математических наук,
профессор,
профессор механико-математического
факультета МГУ имени М.В.Ломоносова

Официальные оппоненты: **Богачев Владимир Игоревич**
доктор физико-математических наук,
профессор,
профессор кафедры теории функций
и функционального анализа
механико-математического факультета
МГУ имени М.В.Ломоносова

Борисов Игорь Семенович
доктор физико-математических наук,
профессор,
главный научный сотрудник
Института математики им. С.Л.Соболева
Сибирского отделения РАН

Жуковский Максим Евгеньевич
доктор физико-математических наук,
доцент,
доцент кафедры дискретной
математики факультета инноваций
и высоких технологий МФТИ

Защита диссертации состоится «26» ноября 2021 г. в 17 ч. 00 мин. на заседании диссертационного совета МГУ.01.07 Московского государственного университета имени М.В.Ломоносова по адресу: 119991, Москва, ГСП-1, Ленинские горы, д. 1, МГУ имени М.В.Ломоносова, Главное здание, механико-математический факультет, ауд. 12-25.

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 сайте ИАС «ИСТИНА»: http://istina.msu.ru/dissertations/397190155

Автореферат разослан «25» октября 2021 г.

Ученый секретарь
диссертационного совета МГУ.01.07,
кандидат физико-математических наук, Раутиан
доцент Надежда Александровна

Актуальность

Диссертация посвящена некоторым методам отбора значимых факторов (признаков), влияющих на изучаемый случайный отклик, а также новым оценкам условной энтропии и взаимной информации в рамках смешанной модели. В таких моделях вектор объясняющих переменных является абсолютно-непрерывным, а переменная отклика имеет дискретное распределение. Эти результаты также важны для идентификации набора значимых факторов.

В настоящее время необходимость применения методов анализа данных и отбора значимых факторов возникает в самых различных областях: от медицины и биологии до финансов и маркетинга[1,2]. С развитием технологий увеличивается и количество данных, которые необходимо анализировать. За последние десятилетия был предложен целый ряд новых методов для решения задач классификации и регрессии[3,4]. Многие из них могут давать более точные прогнозы, если из данных удалить шумовые и избыточные объясняющие переменные. Для этого стали разрабатываться новые методы отбора значимых признаков. Кроме того, отбор значимых признаков помогает создавать интерпретируемые модели, что особо важно в медико-биологических исследованиях, например, для определения генетических факторов, влияющих на возникновение того или иного заболевания.

Методам отбора значимых признаков посвящено огромное количество работ[5,6,7]. Множество стохастических и эвристических методов, направленных на выявление эпистаза (в генетике), рассматривается в работах[8,9].

---

[1] J. Gronsbell, J. Minnier, S. Yu, K. Liao, and T. Cai. Automated feature selection of predictors in electronic medical records data. *Biometrics*, 75(1):268–277, 2019.

[2] L. Lin, D. Liang, C.-C. Yeh, and Huang J.C. Novel feature selection methods to financial distress prediction. *Expert Systems with Applications*, 41:2472–2483, 2014.

[3] Ю. Ю. Линке, И. С. Борисов. О построении явных оценок в задачах нелинейной регрессии. *Теория вероятн. и ее примен.*, 63(1):29—56, 2018.

[4] N. Cristianini and J. Shawe-Taylor. *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods.* Cambridge University Press, 2000.

[5] S.E. Ahmed. *Penalty, Shrinkage and Pretest Strategies. Variable Selection and Estimation.* Springer, Cham, 2014.

[6] P. Bühlmann and S. van de Geer. *Statistics for High-Dimensional Data. Methods, Theory and Applications.* Springer, Heidelberg, 2011.

[7] S. Solorio-Fernández, J.A. Carrasco-Ochoa, and J.F. Martínez-Trinidad. A review of unsupervised feature selection methods. *Artificial Intelligence Review*, 53:907—948, 2020.

[8] A. Bulinski. Some statistical methods in genetics. In V. Schmidt, editor, *Stochastic Geometry, Spatial Statistics and Random Fields. Lecture Notes in Mathematics*, volume 2120, 293–320. Springer-Verlag, Berlin, 2014.

[9] J.H. Moore and S.M. Williams, editors. *Epistasis: Methods and Protocols. Methods in Molecular Biology*, volume 1253. Springer Science + Business Media, New York, 2015.

В главе 1 диссертации предлагается модификация MDR метода (*multifactor-dimensionality reduction*), основанная на использовании стратифицированной выборки вместо выборки из независимых одинаково распределённых наблюдений.

MDR метод впервые был предложен в работе[10] и разработан для обнаружения значимых (в определенном смысле) факторов, существенно влияющих на значение бинарной переменной отклика. Обзор, представленный в работе[11], демонстрирует огромную популярность этого метода. Только в период с 2001 по 2014 год было опубликовано около 800 работ, посвящённых его модификациям, обобщениям и практическим применениям. В работах[12,13,14] был предложен новый подход к введению MDR метода на основе функционала ошибки предсказания функции отклика. Кроме того, в этих статьях был доказан ряд асимптотических свойств возникающих статистических оценок. Однако в случае, если анализируемая выборка из независимых одинаково распределенных наблюдений несбалансирована, т.е. количество наблюдений одного из классов в выборке намного больше, чем другого, то базовый MDR метод может быть недостаточно точен для выборок малого размера, поэтому в нашей работе[15] была предложена модификация метода для статифицированных выборок. В стратифицированной выборке наблюдения уже не являются независимыми в совокупности, поэтому используемая в MDR методе оценка функционала ошибки требует изменений. Для рассматриваемой модификации доказана сильная состоятельность оценки функционала ошибки, а также состоятельность процедуры отбора значимых факторов при условии, что их число известно. Также в первой главе дается новый стоимостной подход для сравнения предложенной модификации с базовым MDR-EFE методом, проведены компьютерные симуляции, продемонстрировавшие преимущества данной модификации.

Глава 2 посвящена новой оценке условной энтропии дискретной слу-

[10]M. D. Ritchie, L. W. Hahn, N. Roodi, R. Bailey, W. D. Dupont, F. F. Parl, and J.H. Moore. Multifactor dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. *Amer. J. Human Genetics*, 69:139–147, 2001.

[11]A. Gola, J.M.M. John, K. van Steen, and R. König. A roadmap to multifactor dimensionality reduction methods. *Briefings in Bioinformatics*, 1–16, 2015.

[12]A. Bulinski, O. Butkovsky, V. Sadovnichy, A. Shashkin, P. Yaskov, A. Balatskiy, Samokhodskaya L., and V. Tkachuk. Statistical methods of snp data analysis and applications. *Open Journal of Statistics*, 2(1):73–87, 2012.

[13]А.В. Булинский. К основам метода понижения размерности объясняющих переменных. *Учреждение Российской академии наук Санкт-Петербургское отделение Математического института им. В.А. Стеклова РАН*, 408:84–101, 2012.

[14]A. Bulinski and A. Rakitko. MDR method for nonbinary response variable. *J. of Multivariate Analysis*, 135:25–42, 2015.

[15]A. Bulinski and A. Kozhevin. New version of the MDR method for stratified samples. *Statistics, Optimization and Information Computing*, 5:1–18, 2017.

чайной величины, принимающей конечное число значений, при условии, задаваемом абсолютно непрерывным вектором. Следует отметить вклад в развитие понятия энтропии, который внесли работы Л.Больцмана, Дж.Гиббса, М.Планка, К.Шеннона, А.Н.Колмогорова, Я.Г.Синая, А.Реньи, К. Цаллиса, А.С.Холево. В различных задачах возникает необходимость использования статистических оценок энтропии, построенных по независимым одинаково распределённым наблюдениям. Например, такие оценки важны при выборе значимых переменных[16] и выявлении неоднородностей в материалах[17], а также применяются в разнообразных областях физики, в задачах теории информации и машинного обучения[18,19]. О прочих задачах, в которых используются статистические оценки энтропии можно прочитать, например, в работе[20]. Был разработан ряд подходов к оцениванию энтропии в различных моделях[21,22,23,24].

Главная цель данной главы — ввести новую статистическую оценку условной энтропии Шеннона для смешанной модели, в которой дискретная переменная отклика со значениями в произвольном конечном множестве зависит от вектора, состоящего из факторов (признаков), имеющих плотность относительно меры Лебега на пространстве $\mathbb{R}^d$. Смешанной моделью является, например, известная модель логистической регрессии[25]. Предлагаемая оценка основана на статистиках k-ближайших соседей, где $k = k_n$ зависит от числа наблюдений n (со статистиками k-ближайших соседей можно ознакомиться, например, в книге[26]). Отметим, что предлагаемая оценка не использует известные статистики Козаченко-Леоненко[27].

---

[16]H. Peng, F. Long, and C. Ding. Feature selection based on mutual information criteria of maxdependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(8):1226–1238, 2005.

[17]P. Alonso-Ruiz and E. Spodarev. Entropy-based inhomogeneity detection in porous media. *arXiv:1611.02241*, 2016.

[18]D. Granziol, B. Ru, S. Zohren, X. Doing, M. Osborne, and S. Roberts. Meme: An accurate maximum entropy method for efficient approximations in large-scale machine learning. *Entropy*, 21(6):1–18, 2019.

[19]R. M. Gray. *Entropy and Information Theory*. Springer US, 2011.

[20]D. Pál, B. Póczos, and C. Szepesvári. Estimation of rényi entropy and mutual information based on generalized nearest-neighbor graphs. In *Proceedings of the 23rd International Conference on Neural Information Processing Systems*, 1849–1857. Vancouver, British Columbia, Canada, 2010.

[21]E. Archer, I.M. Park, and J.W. Pillow. Bayesian entropy estimation for countable discrete distributions. *Journal of Machine Learning Research*, 15:2833–2868, 2014.

[22]I. Nemenman, F. Shafee, and W. Bialek. Entropy and inference, revisited. In *Advances in Neural Information Processing Systems*, 471–478. MIT Press, Cambridge, MA, 2002.

[23]L. Paninski. Estimation of entropy and mutual information. *Neural Computation*, 15:1191–1253, 2003.

[24]Q. Wang, S.R. Kulkarni, and S. Verdú. Divergence estimation for multidimensional densities via k-nearest-neighbor distances. *IEEE transactions on information theory*, 55(5):2392–2405, 2009.

[25]J.M. Hilbe. *Practical Guide to Logistic Regression*. CRC Press, Boca Raton, 2015.

[26]G. Biau and L. Devroye. *Lectures of the Nearest Neighbor Method*. Springer, Cham, 2015.

[27]Л. Ф. Козаченко, Н. Н. Леоненко. О статистической оценке энтропии случайного вектора. *Пробл. передачи информ.*, 23:9–16, 1987.

Для новой оценки устанавливается асимптотическая несмещённость и L_2-состоятельность при стремлении размера выборки к бесконечности. Исследование оценок условной энтропии интересно по следующей причине. Взаимная информация двух случайных величин (или векторов) представляется в виде разности безусловной энтропии одной случайной величины и условной энтропии этой случайной величины при условии второй. Эта информационная характеристика двух случайных величин может быть использована для нахождения значимых факторов, существенно влияющих на значение рассматриваемой переменной отклика. Подобная задача возникает в медицинских и биологических приложениях. Таким образом, статистические оценки взаимной информации, предложенные в диссертации, могут быть использованы для отбора значимых факторов. Это направление исследований развивается в следующей главе.

В главе 3 предлагается оценка взаимной информации абсолютно непрерывного случайного вектора и дискретной случайной величины, принимающей произвольное конечное число значений, основанная на описанной в главе 2 оценке условной энтропии. Взаимная (совместная) информация (*mutual information*) двух случайных векторов широко используется в различных задачах, например, как отмечено выше, для отбора значимых факторов и классификации[28,29]. Для предлагаемой новой оценки и оценки взаимной информации, введенной в работе[30], доказаны асимптотическая несмещённость и L_2-состоятельность. Также приводятся результаты компьютерных симуляций, проведенных для сравнения нескольких известных оценок взаимной информации в рамках смешанной модели. Они продемонстрировали преимущества предложенной новой оценки.

В главе 4 описывается процедура отбора значимых признаков, использующая разработанную оценку взаимной информации для смешанной модели. В предположении, что число значимых признаков известно, доказывается состоятельность описанной процедуры. Также в главе приводятся результаты численных экспериментов, оценивающих точность предлагаемого метода.

Цели и задачи

Цель работы — разработка новых методов идентификации значимых

[28]F. Macedo, R. Oliveira, A. Pacheco, and R. Valadas. Theoretical foundations of forward feature selection methods based on mutual information. *Neurocomputing*, 325:67–89, 2019.

[29]M. Verleysen, F. Rossi, and D. Franëcois. Advances in feature selection with mutual information. *arXiv:0909.0635v1*, 2009.

[30]F. Coelho, A.P. Braga, and M. Verleysen. A mutual information estimator for continuous and discrete variables applied to feature selection and classification problems. *International Journal of Computational Intelligence Systems*, 9(4):726–733, 2016.

признаков, влияющих на изучаемый случайный отклик. А именно, предложено обобщение MDR метода на случай стратифицированной выборки и построены новые оценки взаимной информации и условной энтропии в смешанной модели, когда вектор объясняющих переменных является абсолютно-непрерывным, а переменная отклика имеет дискретное распределение. Установлена состоятельность введенной процедуры отбора значимых признаков.

Научная новизна

В главе 1 вводится модификация оценки MDR-EFE для случая стратифицированной выборки, которая используется для обнаружения значимых факторов среди дискретных случайных величин $X_1, \ldots, X_n$ в модели с дискретной переменной отклика Y. Для данной модифицированной оценки доказывается её сильная состоятельность. Вводится обобщённая XOR-модель. Дается стоимостной подход, позволяющий сравнивать предложенную оценку для стратифицированной выборки с оценкой MDR-EFE для выборки из независимых одинаково распределённых случайных величин.

В главе 2 определяется новая оценка условной энтропии дискретной случайной величины Y при условии случайного вектора X, имеющего абсолютно непрерывное распределение. Оценка основана на использовании статистик ближайших соседей[31]. При широких условиях доказывается асимптотическая несмещённость и L_2-состоятельность упомянутой оценки.

В главе 3 доказывается асимптотическая несмещенность и L_2-состоятельность оценки взаимной информации, введённой в работе[32], а также определяется новая оценка и доказываются ее аналогичные свойства.

В главе 4 на основе предложенной оценки взаимной информации предлагается процедура отбора значимых факторов, и в предположении, что число значимых признаков известно, доказывается ее состоятельность.

Положения, выносимые на защиту

1. Оценка функционала ошибки MDR-EFE метода в случае стратифицированных выборок является сильно состоятельной, а предложенная на ее основе модификация метода позволяет добиться более точных результатов (в смысле стоимостного подхода) по сравнению с базовым MDR методом в случае несбалансированных выборок.

[31] См. сноску 26 выше

[32] См. сноску 30 выше

2. Введенная оценка условной энтропии в смешанной модели является асимптотически несмещенной и L_2-состоятельной при выполнении весьма широких условий.

3. Оценка взаимной информации, построенная на основе предложенной оценки условной энтропии, является асимптотически несмещенной и L_2-состоятельной, а также более точной по сравнению с рядом других оценок в рамках смешанных моделей.

4. Предложенная оценка взаимной информации используется для построения процедуры отбора значимых признаков и при этом позволяет получить состоятельную оценку набора значимых признаков.

Теоретическая и практическая значимость

Работа носит теоретический характер. Ее результаты могут быть использованы для дальнейшего развития теории методов отбора значимых признаков, а также для практического применения при анализе данных наблюдений.

Методология и методы диссертационного исследования

В диссертации использованы методы теории вероятностей и математической статистики, а также математического анализа. В **главе 1** мы применяем закон больших чисел для треугольных массивов и метод кросс-валидации (перекрестной валидации) для оценивания ошибки предсказания. В **главе 2** используются свойства условного математического ожидания, различные виды неравенств концентрации. Существенную роль играет аппарат теории меры[33]. **Главы 3 и 4** основаны на результатах второй главы и используют предельные теоремы для статистических оценок дифференциальной энтропии Шеннона.

Соответствие паспорту научной специальности

В диссертации изучаются различные статистические методы отбора значимых признаков, оценки энтропии в смешанной модели и исследуются их свойства, в силу чего диссертация соответствует паспорту специальности 01.01.05 "Теория вероятностей и математическая статистика".

[33] В.И. Богачев. *Основы теории меры. Т.1,2, 2-е изд.* НИЦ «Регулярная и хаотическая динамика», 2003.

Степень достоверности и апробация результатов

Основные результаты диссертационной работы были доложены на трех международных конференциях и всероссийской конференции:

- 3rd International Workshop «Analysis, Geometry and Probability», Ульм, Германия, 28 сентября - 3 октября 2015 г.
  Тема доклада: Variable selection for overlapping groups.

- 9th International Workshop on Applied Probability (IWAP 2018), Будапешт, Венгрия, 18-21 июня 2018 г.
  Тема доклада: Feature selection and the mutual information estimation.

- 4th International Conference on Stochastic Methods, г. Новороссийск, Россия, 2-9 июня 2019 г.
  Тема доклада: An information theory-based approach to feature selection

- Третья Санкт-Петербургская зимняя молодежная конференция по теории вероятностей и математической физике, Санкт-Петербург, Россия, 16-18 декабря 2019 г.
  Тема доклада: Критерии остановки для процедуры отбора значимых признаков.

По теме диссертации автором были сделаны следующие доклады на научно-исследовательских семинарах:

- Большой семинар кафедры теории вероятностей МГУ под руководством академика РАН, проф. А.Н. Ширяева (МГУ, Москва, 2019 г.),

- «Forschungsseminar Stochastische Geometrie und räumliche Statistik» под руководством Prof. E.Spodarev (Institut für Stochastik, Ulm University, Germany, 2016 г.),

- Аспирантский коллоквиум кафедры теории вероятностей под руководством академика РАН, проф. А.Н. Ширяева (МГУ, Москва, 2018 и 2019 гг.),

- «Современные проблемы фундаментальной математики и механики» под руководством академика РАН, проф. В.А.Садовничего (МГУ, Москва, 2019 г.),

- Городской семинар по теории вероятностей и математической статистике под руководством академика РАН, проф. И.А.Ибрагимова (ПОМИ РАН, Санкт-Петербург, 2019 г.),

- «Асимптотический анализ случайных процессов и полей» под руководством проф. А.В. Булинского (МГУ, Москва, 2015-2020 гг.)

Публикации

Основные результаты диссертации изложены в 7 публикациях автора. Из них 4 статьи (см. [1]-[4], работы [1]-[3] в соавторстве), которые опубликованы в рецензируемых научных журналах, входящих в базы SCOPUS и Web of Science. В материалах международных конференций представлены 3 публикации [5]-[7].

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 1.1, 2.6, 2.8 и 3.9. Автору диссертации принадлежит доказательство теорем, следствий, остальных лемм, а также проведение компьютерных симуляций. В списке из 7 работ по теме диссертации, приведенном на страницах 18 и 19 автореферата, в совместных статьях [1], [2] и [3] указан вклад каждого автора.

Структура диссертации

Диссертация, объемом 129 страниц, состоит из введения, четырех глав, заключения и списка литературы, насчитывающего 106 наименований. В заключении к диссертации сформулированы возможные направления дальнейших исследований.

В диссертацию вошли результаты, полученные при работе по проекту 14-21-00162 Российского научного фонда (руководитель гранта - академик РАН А.Н.Ширяев).

Основное содержание работы

Во введении к диссертации обсуждается актуальность, научная новизна и структура диссертации, формулируются цели и задачи работы, описывается теоретическая, практическая значимость и методы, используемые для получения результатов.

В главе 1 исследуется модификация оценки MDR-EFE функционала ошибки прогноза отклика в случае стратифицированной выборки. В разделе 1.1 приводится постановка задачи. В разделе 1.2 даются необходимые

определения и обозначения, формулируется основной вспомогательный результат первой главы. Пусть $\mathbb{X} = \mathbb{X}_1 \times \cdots \times \mathbb{X}_d$, где для каждого $i = 1, \ldots, d$ множество $\mathbb{X}_i$ — конечное. Пусть $(X^1, Y^1), (X^2, Y^2), \ldots$ – последовательность независимых векторов со значениями в $\mathbb{X} \times \{-1, 1\}$, имеющих то же распределение, что (X, Y). Рассмотрим $(Y^k)_{k\in\mathbb{N}}$ и выберем (случайные) индексы $l \le j_{-1}^1 < j_{-1}^2 < \ldots$, для которых $Y^{j_{-1}^k} = -1$, $k \in \mathbb{N}$. Аналогично все наблюдения Y^i со значением 1 обозначим $Y^{j_1^k}$, где $1 \le j_1^1 < j_1^2 < \ldots$. Рассмотрим выборки

$$\zeta_{n_1}^1 = \{(X^{j_1^1}, 1), \ldots, (X^{j_1^{n_1}}, 1)\}, \;\; \zeta_{n_{-1}}^{-1} = \{(X^{j_{-1}^1}, -1), \ldots, (X^{j_{-1}^{n_{-1}}}, -1)\},$$

где $n_1, n_{-1} \in \mathbb{N}$. Пусть $n := n_1 + n_{-1}$ – объем стратифицированной выборки $\zeta_n = \zeta_{n_1}^1 \cup \zeta_{n_{-1}}^{-1}$. Таким образом, в отличие от выборки $(X^1, Y^1), \ldots, (X^n, Y^n)$, состоящей из независимых одинаково распределенных векторов, имеются две подвыборки $\zeta_{n_1}^1$ и $\zeta_{n_{-1}}^{-1}$, которые, согласно лемме 1.1, состоят из наблюдений, распределенных соответственно, как X при условии $Y = 1$, и X при условии $Y = -1$. Для $y \in \{-1, 1\}$, $K \in \mathbb{N}$ и $k = 1, \ldots, K$,

$$S_k^y(n_y) := \left\{ j_y^i : i \in \left\{ (k-1) \left[\frac{n_y}{K}\right], \ldots, k \left[\frac{n_y}{K}\right] \mathbb{I}\{k < K\} + n_y \mathbb{I}\{k = K\} \right\} \right\}.$$

Будем использовать оценки $\widehat{P}_n^y$ вероятностей $\mathsf{P}(Y = y)$ такие, что $\widehat{P}_n^y \to \mathsf{P}(Y = y)$ п.н. при $n \to \infty$ для $y \in \{-1, 1\}$. Такие оценки, как объяснено в разделе 1.3, нетрудно построить.

Пусть $f_{PA}(x, \zeta_n, \widehat{P}_n^y)$ – функция, задающая алгоритм предсказания отклика, т.е. функция со значениями во множестве $\{-1, 1\}$, определенная для $x \in \mathbb{X}$, ζ_n и $\widehat{P}_n^y$. Точнее говоря, рассматривается семейство функций, для которых вместо ζ_n берутся подвыборки. Поэтому пишем $f_{PA}(x, \zeta_n(S), \widehat{P}_n^y)$, где $\zeta_n(S) := \{(X^j, Y^j), j \in S\}$, $S \subset (\{j_1^1, \ldots, j_1^{n_1}\} \cup \{j_{-1}^1, \ldots, j_{-1}^{n_{-1}}\})$. Для каждой функции $f : \mathbb{X} \to \{-1, 1\}$ находим f_{PA}, которая в некотором смысле близка к f и используется для оценивания $Err(f)$.

Главную роль в MDR-EFE методе играет функционал ошибки функции прогноза отклика $f \colon \mathbb{X} \to \{-1, 1\}$:

$$Err(f) = E|Y - f(X)|\psi(Y),$$

где $\psi \colon \{-1, 1\} \to \mathbb{R}_+$ — некоторая штрафная функция. Оценку для него в случае стратифицированной выборки можно задать следующим образом, воспользовавшись методом кросс-валидации:

$$\widehat{Err}_K(f_{PA}, \zeta_n, \widehat{\mathsf{P}}_n) := \frac{2}{K} \sum_{y\in\{-1,1\}} \sum_{k=1}^{K} \sum_{j\in S_k^y(n_y)} \frac{\widehat{\psi}(y, \zeta_n(\overline{S_k(n)}), \widehat{\mathsf{P}}_n)\mathbb{I}\{f_{PA}^j(n,k) \neq y\}\widehat{\mathsf{P}}_n^y}{\sharp S_k^y(n_y)},$$

где

$$S_k(n) = S_k(n,\omega) := S_k^1(n_1,\omega) \cup S_k^2(n_2,\omega),$$
$$\overline{S_k(n)} := (\{j_1^1 \ldots, j_1^{n_1}\} \cup \{j_{-1}^1, \ldots, j_{-1}^{n_{-1}}\}) \setminus S_k(n),$$

$f_{PA}^j(n,k) = f_{PA}(X^j, \zeta_n(\overline{S_k(n)}), \mathsf{P}_n^y)$, $\widehat{\psi}$ – оценка ψ, $\sharp$ обозначает мощность конечного множества.

Приведем основную теорему главы 1, являющуюся критерием сильной состоятельности введенных оценок.

Теорема 1.3 *Пусть ζ_n — выборка, описанная выше, ψ — некоторая штрафная функция, $f\colon \mathbb{X} \to \{-1,1\}$ — произвольная функция и f_{PA} определяет алгоритм предсказания. Предположим, что для всех $k = 1, \ldots, n$ выполняется условие*

$$\widehat{\psi}(y, \zeta_n(\overline{S_k(n)}), \widehat{\mathsf{P}}_n) \to \psi(y) \text{ п.н., } n \to \infty, \quad y \in \{-1,1\},$$

и существует непустое множество $U \subset \mathbb{X}$ такое, что для всех $x \in U$ и $k = 1, \ldots, K$, соотношение

$$f_{PA}(x, \zeta_n(\overline{S_k(n)}), \widehat{\mathsf{P}}_n) \to f(x) \text{ п.н., } n \to \infty$$

выполняется. Тогда для всех $a \in (0,1)$ ($n_1 = \max\{[an], 1\}$, $n_{-1} = n - n_1$),

$$\widehat{Err}_K(f_{PA}, \zeta_n, \widehat{\mathsf{P}}_n) \to Err(f) \text{ п.н., } n \to \infty$$

тогда и только тогда, когда

$$\sum_{k=1}^{K} \sum_{y \in \{-1,1\}} \sum_{x \in \mathbb{X}_y} y \mathbb{I}\{f_{PA}(x, \zeta_n(\overline{S_k(n)}), \widehat{\mathsf{P}}_n) = -y\} L(x) \to 0 \text{ п.н.}, \quad n \to \infty,$$

где

$$L(x) := \psi(1)\mathsf{P}(X = x, Y = 1) - \psi(-1)\mathsf{P}(X = x, Y = -1), \quad x \in \mathbb{X},$$
$$\mathbb{X}_y = (\mathbb{X} \setminus U) \cap \{x \in \mathbb{X} : f(x) = y\}, \quad y \in \{-1,1\}.$$

Данная теорема позволяет сформулировать вполне естественные условия, при которых оценка функционала ошибки в методе MDR-EFE, основанная на стратифицированной выборке, является строго состоятельной (см. замечания 1.4-1.6 диссертации).

В разделе 1.4 приводится предложенный автором диссертации стоимостной подход к проведению экспериментов и результаты компьютерных

симуляций для сравнения введенной оценки с существующими. Эксперименты показали преимущества предложенного метода по сравнению с его начальной версией.

Главы 2 и 3 посвящены исследованию оценок энтропии и взаимной информации в смешанной модели. Предположим, что X принимает значения в пространстве $\mathbb{R}^d$, снабженном σ-алгеброй $\mathcal{B}(\mathbb{R}^d)$ (таким образом, $S = \mathbb{R}^d$ и $\mathcal{B} = \mathcal{B}(\mathbb{R}^d)$), Y принимает значения во множестве M и $\mathcal{A} = 2^M$, т.е. состоит из всех подмножеств M. Пусть $\mathsf{P}(Y = y) > 0$ для всех $y \in M$. Предположим, что $\mathsf{P}_{X,Y} \ll \mu \otimes \lambda$, где μ — мера Лебега на $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$, λ — считающая мера на $(M, 2^M)$. Такую модель мы будем называть смешанной. Отметим, что описанная модель включает в себя классическую логистическую регрессию[34].

В главе 2 приводится новая оценка энтропии для смешанной модели. В разделе 2.1 формулируется исследуемая модель и приводятся необходимые определения. Раздел 2.2 начинается с описания предлагаемой оценки энтропии, построенной по последовательности независимых одинаково распределенных случайных векторов. Введем оценку $H(Y|X)$ по выборке $Z^1, \dots, Z^n$:

$$\widehat{H}_{n,k} = \frac{1}{n} \sum_{i=1}^{n} \widehat{H}_{n,k,i}.$$

Здесь $Z^i = (X^i, Y^i)$, $i \in \mathbb{N}$, $n \in \mathbb{N}$, $n > 1$, $k = k(n) \in \{1, \dots, n-1\}$,

$$\widehat{H}_{n,k,i} = -\log(\xi_{n,k,i}(Z^1, \dots, Z^n) + 1) + \log k,$$

$$\xi_{n,k,i}(Z^1, \dots, Z^n) := \sharp\{j \in \{1, \dots, n\} \setminus \{i\} \colon Y^j = Y^i,\ \|X^i - X^j\| \leq \|X^i - X^{i,(k)}\|\},$$

где $\sharp$ обозначает мощность конечного множества, $\|\cdot\|$ — евклидова норма в пространстве $\mathbb{R}^d$, $X^{i,(k)}$ является k-ым ближайшим соседом X^i в выборке $\{X^1, \dots, X^n\} \setminus \{X^i\}$. Очевидно, что случайная величина $\xi_{n,k,i}(Z^1, \dots, Z^n)$ принимает значения $0, 1, \dots, k$.

Определение 2.1 *Функция $g \colon \mathbb{R}^d \to \mathbb{R}$ называется локально стягиваемой в точке $x \in \mathbb{R}^d$, если существуют строго положительные $R_0(x)$ и $C_0(x)$ такие, что*

$$\left| g(x) - \frac{1}{|B(x,R)|} \int_{B(x,R)} g(v)\, dv \right| \leq C_0(x) R \quad \text{для} \quad R \in (0, R_0(x)), \qquad (1)$$

[34] L. Massaron and A. Boschetti. *Regression Analysis with Python.* Packt Publishing Ltd., Birmingham, 2016.

где $|B(x,R)|$ *— объём шара* $B(x,R) := \{v \in \mathbb{R}^d : \|v-x\| \le R\}$, *т.е.* $|B(x,R)| = \mu(B(x,R))$. *Функция* g *называется* C_0*-стягиваемой, если она локально стягиваема для* μ*-почти всех точек* $x \in \mathbb{R}^d$ *и, кроме того, для всех* x *выполняется неравенство* $C_0(x) \le C_0$ *и* $R_0(x) \ge R_0$, *где* C_0 *и* R_0 *— строго положительные константы.*

Заметим, что это требование выполнено для любой функции, удовлетворяющей условию Липшица.

Далее приводятся два основных результата второй главы. Предполагается, что множество M конечно и $\mathsf{P}(Y=y)>0$ для всех $y \in M$.

Теорема 2.3 *Пусть в рамках модели, описанной в разделе 2.1, выполняются следующие условия. Для каждого фиксированного* $y \in M$ *и* μ*-почти всех* $x \in \mathbb{R}^d$ *функция плотности случайного вектора* (X,Y) $f(x,y)$, *т.е.* $f(\cdot,y)$, *строго положительна и* C_0*-стягиваема,*

$$k = k_n \propto n^{\alpha}$$

для некоторого $\alpha \in (0,1)$, *и для некоторого* $\varepsilon > 0$

$$\mathsf{E}|\log f_X(X)|^{1+\varepsilon} < \infty, \tag{2}$$

где $f_X(\cdot)$ *— плотность* X. *Тогда*

$$\mathsf{E}\widehat{H}_{n,k} \to H(Y|X), \;\; n \to \infty,$$

т.е. $\widehat{H}_{n,k}$ *является асимптотически несмещённой оценкой* $H(Y|X)$.

Теорема 2.4 *Пусть условие* (2) *теоремы 2.1 заменяется следующим. Для некоторого* $\varepsilon > 0$

$$\mathsf{E}|\log f_X(X)|^{2+\varepsilon} < \infty.$$

Тогда

$$\mathsf{E}(\widehat{H}_{n,k} - H(Y|X))^2 \to 0, \;\; n \to \infty,$$

т.е. $\widehat{H}_{n,k}$ *является* L_2*-состоятельной оценкой* $H(Y|X)$.

Раздел 2.3 содержит несколько лемм, две из которых играют важную роль при доказательстве теорем 2.3 и 2.4. Раздел 2.4 содержит доказательства основных результатов.

Оценкам взаимной информации посвящена **глава 3**. В разделе 3.1 приводится определение оценки взаимной информации $\widehat{I}_{n,k}(X,Y)$ для смешанной модели, предложенной в работе[35], а также устанавливается несколько вспомогательных результатов.

[35] См. сноску 30 выше

Пусть $(X^1, Y^1), (X^2, Y^2), \ldots$ — независимые одинаково распределенные случайные величины такие, что $Law(X^1, Y^1) = Law(X, Y)$, а распределение (X, Y) описывается смешанной моделью, представленной в предыдущей главе. Пусть $\zeta_n = \{(X^1, Y^1), \ldots, (X^n, Y^n)\}$, $n \in \mathbb{N}$. Для $y \in M$ рассмотрим последовательность индексов $1 \leq j_1^y < j_2^y < \ldots$ таких, что $Y^{j_k^y} = y$. Тогда можно написать

$$\zeta_n^y = \{(X^{j_k^y}, y)\}_{k=1}^{N_{y,n}}, \quad N_{y,n} = \max\{k \in \mathbb{N}\colon j_k^y \leq n\}.$$

Лемма 3.1 *Для любого $y \in M$ и всех $n \in \mathbb{N}$ случайные величины $N_{y,n}$, $X^{j_1^y}$, $X^{j_2^y}, \ldots$ независимы. Более того, для каждого $k \in \{0, \ldots, n\}$*

$$\mathsf{P}(N_{y,n} = k) = \binom{n}{k} \mathsf{P}(Y = y)^k \mathsf{P}(Y \neq y)^{n-k}$$

и $\mathsf{P}(X^{j_i^y} \in B) = \mathsf{P}(X \in B | Y = y)$, где $B \in \mathcal{B}(\mathbb{R}^d)$, $i \in \mathbb{N}$ и $y \in M$.

Чтобы сформулировать основной результат, введем некоторые обозначения. Для $x \in \mathbb{R}$ и $r \geq 0$ положим $B(x, r) = \{v \in \mathbb{R}^d\colon \|x - v\| \leq r\}$. Для $r > 0$, $R > 0$, $\varepsilon > 0$, $\nu > 0$ и вероятностной плотности f (относительно меры μ, заданной на $\mathbb{R}^d$), следуя [36], введем функционалы, принимающие значения в $[0, \infty]$:

$$I_f(x, r) := \frac{\int_{B(x,r)} f(z)\, dz}{r^d V_d},$$

$$M_f(x, R) = \sup_{r \in (0,R]} I_f(x, r), \quad m_f(x, R) = \inf_{r \in (0,R]} I_f(x, r),$$

$$L_f(\nu) := \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} |\log \|x - z\||^\nu f(x) f(z) dx\, dz,$$

$$Q_f(\varepsilon, R) := \int_{\mathbb{R}^d} M_f^\varepsilon(x, R) f(x)\, dx,$$

$$T_f(\varepsilon, R) := \int_{\mathbb{R}^d} m_f^{-\varepsilon}(x, R) f(x)\, dx.$$

Здесь V_d - объем единичного шара в $\mathbb{R}^d$. Для $\varepsilon > 0$ положим $Q_f(\varepsilon) := Q_f(\varepsilon, \varepsilon)$, $T_f(\varepsilon) := T_f(\varepsilon, \varepsilon)$.

Лемма 3.3 *Предположим, что для некоторого $\nu > 1$ и $\varepsilon > 0$, выполняются следующие неравенства:*

$$L_f(\nu) < \infty, \;\; Q_f(\varepsilon) < \infty, \;\; T_f(\varepsilon) < \infty,$$

[36] A. Bulinski and D. Dimitrov. Statistical estimation of the Kullback - Leibler divergence. *Mathematics*, 544(9):1–36, 2021.

если положить $f = f_X(\cdot)$ *и* $f = f_{X|Y}(\cdot|y)$, $y \in M$. *Тогда для всех* $k \in \mathbb{N}$

$$\mathsf{E}\widehat{I}_{n,k}(X,Y) \to I(X,Y), \;\; n \to \infty. \tag{3}$$

Эти леммы необходимы для доказательства основных результатов раздела 3.1: теорем 3.4 и 3.5.

Теорема 3.4 *Предположим, что для некоторых* $\nu > 1$ *и* $\varepsilon > 0$ *выполняются следующие неравенства:*

$$L_f(\nu) < \infty, \;\; Q_f(\varepsilon) < \infty, \;\; T_g(\varepsilon) < \infty, \tag{4}$$

если мы положим $f = f_X$ *и* $g(\cdot) = f_{X,Y}(\cdot, y)$, $y \in M$. *Тогда для всех* $k \in \mathbb{N}$ (3) *справедливо.*

Теорема 3.5 *Пусть выполнены условия теоремы 3.4 и* $L_\nu(f) < \infty$ *для некоторого* $\nu > 2$ (*вместо* $\nu > 1$). *Тогда для всех* $k \in \mathbb{N}$

$$\mathsf{E}(\widehat{I}_{n,k}(X;Y) - I(X;Y))^2 \to 0, \;\; n \to \infty.$$

В разделе 3.2 вводится новая оценка взаимной информации $\widehat{I}^{(1)}_{n,k_n}(X,Y)$, основанная на предложенной в главе 2 оценке энтропии, и доказывается ее асимптотическая несмещенность и L_2-состоятельность.

Теорема 3.6 *Пусть для всех* $y \in M$ *функция* $f_{X,Y}(\cdot, y)$ *является* C_0-*стягиваемой и строго положительной. Предположим, что для* $f = f_X$ *и некоторого* $\varepsilon > 0$ *имеем*

$$\mathsf{E}|\log f(X)|^{1+\varepsilon} < \infty. \tag{5}$$

Тогда для всех $\alpha \in (0,1)$ *и* $k_n \propto n^\alpha$ *выполнено следующее соотношение*

$$\mathsf{E}\widehat{I}^{(1)}_{n,k_n}(X,Y) \to I(X,Y), \;\; n \to \infty,$$

т.е. оценка $\widehat{I}^{(1)}_{n,k_n}(X,Y)$ *является асимптотически несмещенной.*

Теорема 3.7 *Пусть условие* (5) *Теоремы 3.6 заменяется следующим. Предположим, что для некоторого* $\varepsilon > 0$

$$\mathsf{E}|\log f(X)|^{2+\varepsilon} < \infty.$$

Тогда для всех $\alpha \in (0,1)$ *и* $k_n \propto n^\alpha$ *выполнено соотношение*

$$\mathsf{E}(\widehat{I}^{(1)}_{n,k_n}(X,Y) - I(X,Y))^2 \to 0, \;\; n \to \infty,$$

т.е. оценка $\widehat{I}^{(1)}_{n,k_n}(X,Y)$ *является* L_2*-состоятельной.*

В разделе 3.4 приводятся результаты компьютерных симуляций, продемонстрировавшие преимущества предложенной оценки по сравнению с существующими. Раздел 3.5 содержит доказательства вспомогательных результатов.

В главе 4 предлагается процедура выбора значимых факторов, основанная на введенной оценке взаимной информации. В разделе 4.1 приводится основное определение.

Определение 4.1 *Набор индексов* $S = \{s_1, \dots s_m\}$ *и набор переменных* $X_S := (X_{s_1}, \dots, X_{s_m})$*, где* $1 \le s_1 < \dots, < s_m \le d$*, называются значимыми, если при каждом* $y \in M$ *и* μ*-почти всех* $x \in \mathbb{R}^d$ *для условных плотностей выполняется соотношение*

$$f_{Y|X}(y|x) = f_{Y|X_S}(y|x_S). \tag{6}$$

Пусть независимые векторы (X^i, Y^i), $i \in \mathbb{N}$, имеют такое же распределение, как вектор (X,Y). Рассмотрим выборку $\zeta_n = \{(X^i, Y^i)\}_{i=1}^n$. Пусть

$$Q_m = \{L := (l_1, \dots, l_m) : 1 \le l_1 < \dots < l_m \le d\},$$

т.е. Q_m — набор всех подмножеств $\{1, \dots, d\}$, содержащих ровно m элементов. Для любого $L \in Q_m$ определим $\zeta_{n,L} = \{(X^i_L, Y^i)\}_{i=1}^n$ и оценим взаимную информацию $I(X_L; Y)$ для каждой выборки $\zeta_{n,L}$. Для этого используем оценку $\widehat{I}^{(1)}_{n,k}(X,Y)$ (в этой главе она обозначается $\widehat{I}_{n,k}(X,Y)$), при построении которой вместо ζ_n берется $\zeta_{n,L}$. Будем писать $\widehat{I}_{n,k,L} := \widehat{I}_{n,k}(X_L; Y)$, где $k = k(n)$ некоторая функция, $k(n) \in \{1, \dots, n-1\}$.

Введем набор случайных множеств

$$\widehat{\mathbb{S}}_{n,k}(\omega) = \arg\max_{L \in Q_m} \widehat{I}_{n,k,L}(\omega).$$

Таким образом, $\widehat{\mathbb{S}}_{n,k}$ — набор таких множеств $\widehat{S}_{n,k}$, что

$$\max_{L \in Q_m} \widehat{I}_{n,k,L} = \widehat{I}_{n,k,\widehat{S}_{n,k}}$$

для каждого $\widehat{S}_{n,k} \in \widehat{\mathbb{S}}_{n,k}$.

Основной результат формулируется в разделе 4.2 следующим образом.

Теорема 4.2 *Предположим, что для некоторого* $m \in \{1, \dots, n-1\}$ *существует непустое множество* $\mathbb{S}_m$*, состоящее из всех наборов значимых индексов* S *мощности* $|S| = m$*. Пусть имеется строго положительная версия плотности* $f_{X,Y}$*. Пусть также для всех* $L \in Q_m$ *и* $y \in M$

плотность $f_{X_L,Y}(\cdot, y)$ *является* C_0*-стягиваемой и* $\mathsf{E}|\log f_{X_L}(X_L)|^{2+\varepsilon} < \infty$ *для некоторого* $\varepsilon > 0$*. Тогда для всех* $\alpha \in (0,1)$ *и* $k = k(n) \propto n^\alpha$*,*

$$\mathsf{P}(\widehat{\mathbb{S}}_{n,k} \subset \mathbb{S}_m) \to 1 \ \ \textit{при} \ \ n \to \infty.$$

В частности, если набор $\mathbb{S}_m$ *состоит из единственного множества* S_m*, то* $\mathsf{P}(\widehat{\mathbb{S}}_{n,k} = S_m) \to 1$*,* $n \to \infty$*.*

В разделе 4.3 приводятся результаты численных экспериментов, которые показывают, что даже для довольно малых значений k и умеренно больших n предложенная процедура позволяет точно идентифицировать значимые факторы.

Заключение

В диссертационной работе предложена модификация оценки MDR-EFE для стратифицированной выборки в модели с дискретными предикторами и дискретной переменной отклика, доказана её сильная состоятельность. Введен новый стоимостной подход, позволяющий сравнивать предложенную оценку для стратифицированной выборки с оценкой MDR-EFE для выборки из независимых одинаково распределённых случайных величин. Приведены результаты численных экспериментов по сравнению оценок в рамках предложенной обобщенной XOR-модели, продемонстрировавшие преимущества новой оценки.

Введена оценка условной энтропии дискретной случайной величины при условии случайного вектора, имеющего абсолютно непрерывное распределение. Доказана её асимптотическая несмещённость и L_2-состоятельность.

Аналогичные свойства доказаны для оценки взаимной информации, введённой в работе [37], а также для оценки взаимной информации, построенной на основе оценки условной энтропии, предложенной в данной диссертационной работе.

На основе предложенных оценок условной энтропии и взаимной информациии разработана процедура отбора значимых факторов в предположении, что число значимых признаков известно, доказана состоятельность этой процедуры.

Таким образом, полученные в рамках диссертации результаты могут быть использованы для дальнейших исследований в области отбора значимых признаков и применены для решения практических задач, например, для обнаружения факторов, влияющих на возникновение некоторого заболевания.

[37] См. сноску 30 выше

В дальнейшем представляло бы интерес развитие изложенных методов для случая, когда количество значимых факторов заранее неизвестно. В настоящий момент в научной литературе этот вопрос мало изучен[38].

Благодарности

Работа выполнена под научным руководством профессора Александра Вадимовича Булинского, которому автор выражает искреннюю благодарность.

Работы автора по теме диссертации

Статьи в научных журналах Web of Science, SCOPUS, RSCI

[1] A. Bulinski and A. Kozhevin. New version of the MDR method for stratified samples. *Statistics, Optimization and Information Computing*, 5:1–18, 2017.

ИФ SJR - 0.3 / 1.09 п.л. / вклад соискателя 0.9 п.л.

А.В.Булинскому принадлежит постановка задач и общий подход к их решению, им также доказана лемма 1. Все остальные результаты доказаны А.А.Кожевиным, им проведены все компьютерные симуляции.

[2] A. Bulinski and A. Kozhevin. Statistical estimation of conditional Shannon entropy. *ESAIM: Probability and Statistics*, 23:350-386, 2019.

ИФ WoS (JIF) - 0.745 / 2.28 п.л. / вклад соискателя 1.47 п.л.

А.В.Булинскому принадлежит постановка задач и общий подход к их решению, им также доказаны леммы 3.1 и 3.3, все остальные результаты доказаны А.А.Кожевиным.

[3] A. Bulinski and A. Kozhevin. Statistical Estimation of Mutual Information for Mixed Model. *Methodology and Computing in Applied Probability*, 23:123–142, 2021.

ИФ WoS (JIF) - 1.147 / 1.25 п.л. / вклад соискателя 1 п.л.

А.В.Булинскому принадлежит постановка задач и общий подход к их решению, им также доказана лемма 1. Все остальные результаты доказаны А.А.Кожевиным, им проведены все компьютерные симуляции.

[38] См. сноску 30 выше

[4] A. Kozhevin. Feature selection based on estimation of mutual information *Siberian Electronic Mathematical Reports*, 18(1):720-728, 2021.

ИФ WoS (JCI) - 0.45 / 0.56 п.л. / вклад соискателя 0.56 п.л.

Тезисы конференций

[5] A. Kozhevin. Variable selection method with group overlapping. In *3rd International Workshop "Analysis, Geometry and Probability Ulm, Germany*, p.47, 2015.

[6] A. Kozhevin. Feature selection and the mutual information estimation. In *Abstracts of the 9-th International Workshop on Applied Probability 18-21 June 2018, Budapest, Hungary*, p.137–138., 2018.

[7] А. Кожевин. Информационный подход к отбору значимых признаков. *Теория вероятностей и ее применения*, 65(1):171-172, 2020 // A. Kozhevin. An information theory-based approach to feature selection. *Theory of Probability and its Applications*, 65(1):138-139, 2020.