ОТЗЫВ официального оппонента

на диссертацию на соискание ученой степени кандидата физико-математических наук Ракитько Александра Сергеевича на тему: «Идентификация значимых факторов с помощью функционала ошибки», по специальности 1.1.4 – теория вероятностей и математическая статистика

Существует множество ключевых моделей, в которых изучаемая переменная (отклик) Y зависит от некоторого набора переменных (факторов) $X = (X_1, \ldots, X_n)$. В стохастических моделях, как правило, совместное распределение X и Y является неизвестным. Начиная с классических работ Лагранжа и Гаусса, исследователи стремились в заданном классе функций найти такую, которая позволяла бы в некотором смысле «хорошо» аппроксимировать значение Y с помощью всего вектора X. Однако часто оказывается, что отклик зависит лишь от сравнительно небольшой части компонент X. Нетривиальная задача состоит в идентификации этих компонент по набору независимых наблюдений $(X^1, Y^1), \ldots, (X^N, Y^N)$, где все (X^j, Y^j) имеют такое же распределение, как (X, Y). Нахождение этой части компонент X, образующих так называемые «значимые факторы», позволяет строить модели, допускающие не только привлекательную интерпретацию, но и позволяющие существенно ускорить обработку экспериментальных данных. В 1996 году T.Tibshirani в рамках линейного регрессионного анализа ввел метод LASSO (least absolute shrinkage and selection operator), обеспечивающий упомянутую идентификацию. Такой подход стал разрабатываться рядом авторов в геофизических исследованиях, начиная с 1986 года. Сама же идея регуляризации, используемая в этом подходе, восходит к трудам академика А.Н.Тихонова. Отметим, что существует целый ряд областей, в которых предположение о линейном характере зависимости X и Y не является адекватным. В этой связи можно указать на медико-биологические исследования. В 2001 году была опубликована статья M.D.Ritchie и ее соавторов, в которой был предложен метод MDR (multifactor dimensionality reduction), нацеленный на выявление значимых факторов в нелинейных непараметрических моделях. Точнее говоря, используя процедуру кросс-валидации, строился определенный алгоритм, позволяющий выбрать значимые факторы из набора X. Подчеркнем, что, в упомянутой работе и целом ряде других не доказывались теоремы, демонстрирующие границы применимости предложенных алгоритмов. Диссертация А.С.Ракитько посвящена выявлению значимых факторов, влияющих на изучаемый случайный отклик, с помощью статистических оценок функционала, описывающего ошибку его предсказания по части компонент вектора X. Такой метод, введенный А.В.Булинским с соавторами в 2012 году для анализа бинарного отклика, получил название

MDR-EFE (error function estimation). Выполненное автором диссертации исследование по развитию метода MDR-EFE, несомненно, актуально, поскольку доказанные теоремы раскрывают границы применения этого метода для небинарного отклика с любым конечным набором значений.

Рассматриваемая диссертация объемом 110 страниц состоит из введения, трех глав, заключения и списка литературы из 100 работ.

Во введении заявлены цели исследования, описана структура диссертации и дан обзор предшествующих результатов.

В первой главе рассматривается обобщение MDR-EFE метода на не бинарный отклик с любым конечным числом значений. При этом факторы тоже принимают значения в произвольном конечном множестве. Изучение не только бинарного отклика представляется важным. Например, для приложений в медицине это позволяет характеризовать состояние пациента не только как «здоров» или «болен», но дает возможность проводить более детальное описание. Такое обобщение потребовало значительных усилий в преодолении возниккающих аналитических трудностей. Теорема 1 заслуживает внимания, поскольку содержит необходимые и достаточные условия сильной состоятельности введенных в диссертации статистических оценок функционала ошибки прогноза отклика. В теореме 3 на основе теоремы 1 демонстрируется возможность идентификации набора значимых факторов. Установление сильной состоятельности оценок, а не только сходимости по вероятности, является существенным, так как позволяет сопоставлять оценки функционалов ошибки для разных наборов компонент X на событии вероятности единица (и тем самым избегаются поправки Бонферрони). Существенную роль в доказательстве играет использование усиленного закона больших чисел в схеме серий исследуемых случайных величин. Отметим также важную теорему 4, дающую обоснование предложенному методу идентификации значимых факторов, когда вектор факторов не является дискретной случайной величиной, а имеет плотность вероятностей по мере Лебега. Доказательство является весьма сложным. Автору потребовалось построить вспомогательные величины, образующие мартингал, применить для них неравенство Азума (являющееся аналогом неравенства Хёфдинга для независимых слагаемых), использовать понятие сходимости вполне для последовательности случайных величин. К заслугам автора следует отнести и проверку того, что налагаемые условия выполняются для широко используемой модели логистической регрессии.

Во второй главе изучаются асимптотические свойства статистических оценок функционала ошибки. Теорема 7 дает широкие условия справедливости центральной предельной теоремы для построенных оценок. Следует отметить, что для установления этого трудного и глубокого результата потребовалось произвести тонкую регуляризацию упомянутых оценок. При этом явно найдены параметры предельного нормального распределения, а доказательство охватывает небинарный случайный отклик. В связи

с исследованием функционала ошибки А.С. Ракитько обратился также к теории перестановочных величин. При этом полученные им в этой области результаты представляют самостоятельный интерес. Автору диссертации удалось доказать новый вариант центральной предельной теоремы для перестановочных величин (лемма 5) и применить его (теорема 12) для статистических оценок функционала ошибки. При этом явно найдены параметры предельного гауссовского закона. А.С.Ракитько установил также для перестановочных величин (теорема 10) аналог известной теоремы Эрдёша – Каца о максимуме нормированных сумм независимых слагаемых.

Третья глава посвящена новому варианту MDR-EFE метода, относящемуся к последовательному отбору значимых факторов по одному на каждом шаге. Последовательный отбор факторов на основе корреляций или информационных характеристик применялся и ранее, начиная, по-видимому, с работы H.Peng et al. (2005). Такой подход существенно упрощает вычислительные процедуры. Однако весьма трудно обосновать, что он (с большой вероятностью) приведет к набору значимых факторов. Для модели наивного байесовского классификатора теорема 14 впервые дает оценку снизу для вероятности последовательного отбора значимых факторов с помощью MDR-EFE метода. Интересно, что А.С.Ракитько сумел связать изучаемую им задачу с рассмотрением логистической регрессии. Приятно отметить, что теоретические результаты, установленные в диссертации, иллюстрируются примерами компьютерного моделирования.

В заключении автор подводит итог проделанной работы и намечает направления дальнейших возможных исследований по каждой из трех глав.

Учитывая вышеизложенное, можно сказать, что диссертация А.С. Ракитько представляет собой цельное математическое исследование, выполненное на очень высоком научном уровне. Доказанные результаты представляют не только теоретический научный интерес, но также допускают приложения к анализу реальных данных. Установленные теоремы и леммы излагаются с верными, полными доказательствами. Автор диссертации проявил творческие способности при решении сложных и актуальных задач современной математической статистики. Приятно также отметить его эрудицию и владение разнообразной вероятностно-аналитической техникой. Основные результаты диссертации опубликованы в десяти работах автора и прошли всестороннюю апробацию. В начале каждой главы указан вклад А.С.Ракитько в решение поставленных задач при наличии совместных публикаций. Установленные результаты докладывались на 10 международных конференциях. В автореферате подробно и правильно излагается содержание диссертации.

Диссертация аккуратно оформлена, существенных дефектов изложения не замечено, но все же есть небольшое количество замечаний чаще всего редакторского типа. Не аккуратно сформулировано Следствие 5. Было бы желательно в главе 3 рассмотреть не только модель наивного байесовского

классификатора. В связи с доказанным в диссертации обобщением теоремы Эрдёша – Каца на перестановочные величины естественно встает вопрос о получении в дальнейшем новых функциональных предельных теорем в схеме перестановочных величин. В частности, можно ли ослабить в лемме 5 моментные условия?

Вместе с тем, указанные замечания не умаляют значимости выполненного исследования. Диєсертация отвечает требованиям, установленным Московским государственным университетом имени М.В.Ломоносова к работам подобного рода. Содержание диссертации соответствует специальности 1.1.4. «Теория вероятностей и математическая статистика» (по физико-математическим наукам), направления исследований: «Непараметрическая статистика» и «Анализ статистических данных». Диссертация удовлетворяет критериям, определенным пп. 2.1-2.5 Положения о присуждении ученых степеней в Московском государственном университете имени М.В. Ломоносова, а также оформлена согласно требованиям Положения о совете по защите диссертаций на соискание ученой степени кандидата наук, на соискание ученой степени доктора наук Московского государственного университета имени М.В.Ломоносова. Таким образом, соискатель Александр Сергеевич Ракитько заслуживает присуждения ученой степени кандидата физико-математических наук по специальности 1.1.4. «Теория вероятностей и математическая статистика».

Официальный оппонент: доктор физико-математических наук, профессор, профессор кафедры математической статистики факультета вычислительной математики и кибернетики Федерального государственного бюджетного образовательного учреждения высшего образования «Московский государственный университет имени М.В.Ломоносова»

УЛЬЯНОВ Владимир Васильевич

Контактные данные: тел.: 8(495)939-53-94, e-mail: vulyanov@cs.msu.su
Специальность, по которой официальным оппонентом защищена диссертация: 01.01.05 – «теория вероятностей и математическая статистика»

Адрес места работы:
119991, г. Москва, Ленинские горы, МГУ имени М.В. Ломоносова, 2-й учебный корпус, факультет ВМК
Тел.: 8(495)939-53-94; e-mail: vulyanov@cs.msu.su

Подпись сотрудника факультета вычислительной математики и кибернетики МГУ имени М.В.Ломоносова В. В. Ульянова удостоверяю:

Подпись удостоверяю
Ведущий специалист по кадрам
Т.Г. Коваленко