

ОТЗЫВ ОФИЦИАЛЬНОГО ОППОНЕНТА

на диссертацию Стрижова Вадима Викторовича

«Порождение и выбор моделей в задачах регрессии и классификации»,

представленную на соискание ученой степени доктора физико-математических

наук по специальности 05.13.17 – теоретические основы информатики

Диссертационная работа В.В. Стрижова посвящена исследованию одной фундаментальной проблемы математической статистики, проблемы автоматического выбора модели, адекватно описывающей исходный массив эмпирических данных и обладающей при этом минимальной сложностью. Данная проблема имеет непосредственную связь с несколькими разделами современной прикладной математики и теоретической информатики, такими как: теория оценивания обобщающей способности алгоритмов обучения, теория регуляризации и теория вычислительной сложности. Ее исследованию посвящены многочисленные работы отечественных и зарубежных ученых (А.Г. Ивахненко, П.С. Краснощекова, К. Бишоп, Й. Ле Кана, Д. Мак-Кея и др.)

Один из актуальных аспектов исследуемой проблемы связан с повышением интерпретируемости получаемых регрессионных моделей в терминах предметной области. В диссертации предлагается производить поиск подходящей модели в счетном, последовательно порождаемом семействе, обеспечивающем в асимптотике высокую заданную точность. Порождающие элементы (и операции над ними) задаются экспертами с точностью до конечного набора определяемых параметров, что обеспечивает возможность апостериорного оценивания информативности элементов суперпозиции с последующей оптимизацией ее структуры. Для этого применяется двухуровневый байесовский вывод: оценивается значение гиперпараметров модели из заданного параметрического семейства, позволяющее оценить правдоподобие параметров модели и информативность элементов суперпозиции. Проводится анализ пространства параметров и гиперпараметров в условиях мультиколлинеарности. Решается задача оценки ковариационных матриц параметров и гиперпараметров и повышения устойчивости модели путем отбора шумовых объектов и мультикоррелирующих признаков на основе полученных оценок. Таким образом, **актуальность проведенного в диссертации исследования**, посвященного методам построения легко интерпретируемых регрессионных моделей минимальной сложности, **представляется бесспорной.**

В работе предложен новый подход для решения задачи автоматического последовательного выбора моделей регрессионного анализа и классификации. Разработаны методы индуктивного порождения моделей – допустимых суперпозиций элементарных функций из заданного множества, и упрощения порожденных моделей, позволяющие выбирать адекватные модели из некоторого множества допустимых моделей-претендентов. Таким образом, снимается необходимость сравнения большого количества моделей.

Предлагаемый подход является новым, а форма представления материала работы полученные выводы – достоверными. Научные положения, выводы и рекомендации, сформулированные в диссертации, являются обоснованными в достаточной степени. Предложенные автором диссертации решения аргументированы и оценены по сравнению с другими известными решениями.

Диссертационная работа состоит из введения, шести глав и заключения.

Во введении обосновывается актуальность работы, определяются цели и задачи исследования.

В первой главе приводится постановка решаемой задачи. Вводятся основные определения: выборка, модель, функция регрессии. Перечислены основные семейства распределений, используемые при фиксации гипотезы порождения данных, виды функций ошибки. Рассмотрена связь между гипотезой порождения данных и функцией ошибки. Приведена классификация задач регрессии. Описаны способы оценки параметров модели и широко используемые ограничения на множество моделей. С точки зрения рецензента, основными полученными в этой главе результатами, являются методы оптимизации функции ошибки полученной с помощью связанного байесовского вывода.

Вторая глава рассмотрена задача порождения моделей. Необходимость порождения моделей объясняется недостаточностью измеряемых признаков для качественного описания зависимой переменной. В главе перечислены некоторые распространенные методы порождения моделей и описан предлагаемый метод, основанный на прогнозировании структуры модели. В рамках предлагаемого метода модели порождаются последовательно как суперпозиции элементарных функций, заданных экспертами, с набором переменных и констант. Каждая модель, полученная таким образом, оценивается в соответствии с функцией ошибки. Лучшие модели затем модифицируются на основе значений гиперпараметров: гиперпараметры определяют веса элементов суперпозиции, которые затем могут быть заменены другими суперпозициями. Вводится понятие структурной сложности, позволяющее ограничить число получаемых моделей. Выводятся оценки количества моделей ограниченной сложности. Описаны модификация метода,

позволяющая порождать только допустимые суперпозиции и метод поиска изоморфных суперпозиций для упрощения построенных моделей. Предлагается метод прогнозирования структуры модели. В рамках предлагаемого метода предложен способ представления моделей в виде матриц переходов, позволяющих однозначно восстановить дерево суперпозиции модели. По набору прецедентов оцениваются вероятности переходов между элементами суперпозиций и прогнозирование структуры сводится к максимизации дерева переходов. Основным результатом, полученным в этой главе, является метод анализа ковариационных матриц параметров моделей – многомерных случайных величин. На основе этого анализа принимается решение о важности элементов моделей. Основными результатами, полученными в этой главе, являются условия допустимости суперпозиций и метод восстановления структур суперпозиций по прецедентам.

В третьей главе описаны подходы к выбору признаков (в обобщенно-линейных моделях или элементов суперпозиций – в линейных). Перечислены наиболее часто используемые способы выбора признаков. Рассмотрены способы задания расстояния между моделями и условия сходимости моделей к оптимальной при последовательном порождении. Предлагается метод последовательного выбора признаков, построенный таким образом, что добавление очередного признака увеличивает правдоподобие модели, а удаление снижает число мультикоррелирующих признаков. Отбор производится в два этапа: добавление и удаление признаков в соответствии с критерием мультиколлинеарности Белсли. При фиксированной сложности модели функция ошибки рассмотрена как случайная величина. Показано, что функция ошибки достигает минимума на том же наборе признаков, на котором ее дисперсия минимальна. Основным результатом, полученным в этой главе, является метод анализа ковариационных матриц параметров моделей – многомерных случайных величин. На основе этого анализа принимается решение о важности элементов моделей.

В четвертой главе описано использование связанного байесовского вывода при выборе модели. Предложены аналитические и сэмплирующие методы оценки гиперпараметров модели. Продемонстрировано использование получаемых оценок для отбора шумовых и коррелирующих признаков. Ставится задача многоуровневой регрессии. Предложен метод решения задачи одновременного выбора признаков и объектов в задаче на основе параллельно оцениваемых ковариационных матриц параметров и гиперпараметров. Предложен ряд способов получения оценок гиперпараметров моделей.

В пятой главе рассмотрена задача выбора модели для случаев, когда признаковое описание задано в ранговых шкалах. Для приведения объектов к единой линейной шкале решается задача построения интегральных индикаторов объектов. Описаны известные

методы получения интегральных индикаторов. В отличие от ранее использованных методов, предлагаемый метод опирается на экспертные оценки качества объектов. Экспертные оценки описываются конусами в пространстве весов показателей и в пространстве интегральных индикаторов. Проверяется непротиворечивость экспертных оценок, в случае обнаружения противоречивости выполняется процедура уточнения экспертных оценок. Приводится теоретическое обоснование предложенного метода. Предлагаются способы повышения устойчивости индикаторов. Центральным результатом шестой главы, с точки зрения рецензента, является постановка и решение задачи согласования экспертных оценок в ранговых шкалах, сформулированные в терминах задачи регрессионного анализа.

В шестой главе рассмотрено несколько прикладных задач, в рамках которых необходимо решение теоретических задач, поставленных в первой главе. При рассмотрении задач уделяется особое внимание этапам постановки задачи регрессии, описанным в работе: выбору гипотезы порождения данных и функции ошибки, фиксации ограничений на модель регрессии, постановке оптимизационной задачи при настройке параметров. Выделяются ключевые аспекты планирования прикладных исследований. Важным результатом этой главы является решение задачи оптимального управления, поставленной для случая восстановления системы одновременных уравнений в векторной авторегрессии.

В диссертационной работе В.В. Стрижова получены следующие основные результаты.

- Разработаны новые способы построения моделей как суперпозиций параметрических функций из экспертно-задаваемого множества для решения задач регрессии и классификации.
- Предложен ряд новых алгоритмов получения оценок параметров и гиперпараметров регрессионных моделей, основанных на свойствах связанного байесовского вывода.
- Предложен метод последовательного порождения и выбора моделей субоптимальной сложности.
- Исследована проблема мультиколлинеарности при выборе моделей из множества большой мощности. Предложены способы построения устойчивых моделей.
- Предложены методы построения моделей для выборок, элементы которых принимают значения из множеств с заданными на них отношениями полного или частичного порядка.

- Решена обратная задача управления: предложен метод управления фазовой траекторией, описанной системой линейных уравнений специального вида.

По сути представленных результатов замечаний нет. В качестве пожелания к дальнейшим исследованиям хотелось бы отметить следующее. Большая часть полученных в работе результатов получена в предположении о нормальности законов распределения параметров (гиперпараметров) и помех. Интересна возможность распространения этих результатов на случай более общих параметрических систем распределений.

Результаты диссертационной работы изложены в двадцати восьми статьях, опубликованных в журналах, входящих в перечень ВАК РФ, и докладывались на российских и международных конференциях.

Автореферат адекватно отражает содержание диссертационной работы.

Диссертация соответствует критериям, установленным Положением о присуждении ученых степеней утвержденного Постановлением Правительства РФ от 24 сентября 2013 года № 842: автором исследований разработаны теоретические положения, совокупность которых можно квалифицировать как научное достижение в области теории методов анализа данных. Диссертация обладает внутренним единством, содержит новые научные результаты и рекомендации по их практическому использованию.

Работа полностью соответствует требованиям ВАК РФ, предъявляемым к диссертациям на соискание ученой степени доктора физико-математических наук по специальности 05.13.17 — теоретические основы информатики, а ее автор, **В.В. Стрижов**, заслуживает присуждения ему ученой степени доктора физико-математических наук по данной специальности.

Официальный оппонент

Доктор физико-математических наук, с.н.с, заведующий Отделом математического программирования Федерального государственного бюджетного учреждения науки Институт математики и механики им. Н. Н. Красовского УрО РАН

Адрес: 620219, г.Екатеринбург, ГСП-384, ул. Софьи Ковалевской, 16

Тел.: 8 (343) 375 35 05

E-mail: mkhachay@imm.uran.ru

28.10.2014

Хачай Михаил Юрьевич

Подпись заверяю
ченый секретарь
УрО РАН



Ульянов О.Н.