ОТЗЫВ
официального оппонента д.ф.-м.н. С.Д. Двоенко
на диссертацию Ишкиной Шауры Хабировны
«Комбинаторные оценки переобучения пороговых решающих правил»,
представленную на соискание ученой степени
кандидата физико-математических наук по специальности
1.2.1 - «Искусственный интеллект и машинное обучение»

Актуальность темы диссертации. При решении задачи обучения строится алгоритм, который восстанавливает зависимость выходных переменных от входных на объектах из обучающей выборки. В задаче классификации выходная переменная принимает дискретные значения, а алгоритмы называются классификаторами.

Характеристикой всякого хорошего классификатора является его высокая обобщающая способность, т.е. его хорошее качество классификации на новых объектах, ранее не входивших в обучающую выборку. Если же качество классификатора на независимой контрольной выборке оказывается хуже, чем на обучающей выборке, то возникает известный эффект т.н. переобучения.

Получение оценок обобщающей способности семейства решающих правил, в том числе и классификаторов, на основе информации об обучающей выборке и структуре семейства правил до сих пор остается одной из основных задач теории статистического обучения.

В 70-х гг. прошлого века советские ученые В.Н. Вапник и А.Я. Червоненкис сформулировали основные статистические проблемы обучения в терминах задачи минимизации среднего риска, т.е. вероятности ошибки классификатора на новом объекте, предложив методы оценки среднего риска по эмпирическим данным. Были получены равномерные по семействам классификаторов оценки, связывающие вероятность уклонения среднего риска от эмпирического с объёмом обучающей выборки и сложностью (емкостью) семейства решающих правил, для которых минимизируется средний риск. Этот фундаментальный результат активно используется и сегодня.

Проблема оценок Вапника–Червоненкиса в том, что они сильно завышены, поскольку являются оценками наихудшего случая. Поэтому теория статистического обучения продолжает активно развиваться в направлении повышения точности оценок с учетом особенностей данных и конкретных алгоритмов распознавания.

Эта проблема имеет важное практическое значение, т.к. завышенность оценок может приводить к неоптимальному выбору структурных параметров класса решающих правил. Кроме того, завышенные оценки не дают возможности исследовать явление переобучения, оценивать и контролировать степень его проявления при решении реальных задач.

В диссертации рассмотрена проблема поиска верхних оценок обобщающей способности одномерных пороговых решающих правил, а

также их применение для повышения обобщающей способности решающих деревьев.

Основные результаты и их научная новизна. Одним из актуальных направлений развития теории обучения в области обобщающей способности решающих правил является их адаптация к особенностям обучающих данных для получения более тонких оценок переобучения. Основным инструментом для получения оценок переобучения являются специальные функционалы вероятности переобучения и полного скользящего контроля.

Следует отметить, что на практике точность эмпирических оценок функционалов обобщающей способности, как правило, оценивается статистически методом Монте–Карло и зависит от числа случайных разбиений. Вычисление таких оценок по определению требует экспоненциального по общему количеству объектов перебора всех возможных разбиений. При этом лишь для некоторых семейств классификаторов удается аналитически вычислить достигаемые верхние оценки вероятности переобучения.

В предшествующих работах научного руководителя диссертанта д.ф.-м.н. К.В. Воронцова была предложена комбинаторная теория переобучения, позволившая получать точные (достигаемые) оценки функционала вероятности переобучения. Этот функционал определяется как доля разбиений конечного множества объектов на обучающую и контрольную выборки фиксированного объёма, при которых происходит переобучение.

В диссертационной работе Ш.Х. Ишкиной в рамках комбинаторной теории переобучения получены достигаемые верхние оценки для трёх функционалов обобщающей способности одномерных пороговых решающих правил — вероятности переобучения, полного скользящего контроля и ожидаемой переобученности.

В комбинаторной теории оценки вероятности переобучения ранее была известна специальная оценка расслоения–связности, учитывающая особенности модели классификациии обучающей выборки, а также эффекты расслоения и связности. Показано, что благодаря расслоению, классификаторы с высокой вероятностью ошибки вносят малый вклад в переобучение, а благодаря связности, у классификаторов с близкими векторами ошибок также снижается вклад в переобучение. Были получены условия, при которых оценка расслоения–связности является точной.

В частности, оказалось, что такой оценке удовлетворяют т.н. монотонные и унимодальные цепи классификаторов. В практических задачах статистического обучения такие цепи могут порождаться элементарными пороговыми правилами, которые применяются в таких распространенных алгоритмах классификации, как решающие деревья, логические закономерности, алгоритмы вычисления оценок, а также при построении линейных классификаторов методом покоординатной оптимизации. Но при этом делается предположение о существовании безошибочного правила, практически не выполнимое в реальных задачах.

В общем случае пороговые правила порождают семейства классификаторов, которые в данном подходе названы прямыми последовательностями. Ранее для них были получены верхние оценки ожидаемой частоты ошибок на контрольной выборке лишь в частных случаях, где некоторые известные уточнения самих оценок расслоения–связности также давали завышенные верхние оценки ожидаемой частоты ошибок для прямых последовательностей. Проблема завышения таких верхних оценок в настоящий момент оставалась малоизученной.

Таким образом, полученные в диссертационной работе комбинаторные оценки обобщающей способности одномерных пороговых решающих правил и теоремы о представлении оценок обобщающей способности для произвольного семейства являются актуальными и новыми.

В данной работе также предложен новый алгоритм построения дерева решений, в котором в качестве критерия выбора атрибута для разделения узла дерева решений используются достигаемые верхние оценки полного скользящего контроля или ожидаемой переобученности пороговых решающих правил.

Построена т.н. суррогатная модель для быстрого вычисления приближенных оценок обобщающей способности семейства пороговых решающих правил с высокой точностью.

Содержание работы. Диссертационная работа состоит из введения, четырех глав, заключения, списка литературы и приложения. Во введении автор обосновывает актуальность темы диссертации, новизну темы исследования, приводит известные постановки основных задач поиска оценок вероятности переобучения.

В первой главе в рамках комбинаторной теории переобучения введено понятие финитного метода обучения, для которого в случае произвольного семейства классификаторов доказаны теоремы о представлении достигаемых верхних оценок в виде произведения числа разбиений двух непересекающихся множеств объектов генеральной совокупности. Доказано, что свойством финитности обладают рассматриваемые в данной работе методы минимизации эмпирического риска и максимизации переобучения. Показано, что уровень переобучения такого семейства зависит от формы графика числа ошибок при варьировании порогового значения. Доказаны теоремы о верхних оценках обобщающей способности семейства одномерных пороговых решающих правил. Доказана полиномиальная вычислительная сложность алгоритма вычисления достигаемых оценок данного семейства.

Во второй главе исследуется проблема завышенности известных оценок обобщающей способности для пороговых классификаторов по сравнению с достигаемыми верхними оценками, рассчитанными в первой главе. Оценки вероятности переобучения (классические на основе размерности Вапника–

Червоненкиса, комбинаторные оценки расслоения–связности и Е.А.Соколова) и оценки ожидаемой переобученности на основе Радемахеровской сложности оказываются завышены на 1-2 порядка. Показано, что оценки И.С. Гуза для полного скользящего контроля хорошо согласуются с точными оценками, откуда следует вывод о применимости данных оценок в прикладных задачах в частных случаях.

В третьей главе разработанные комбинаторные оценки применены в прикладной задаче планирования трассерных исследований нефтегазовых месторождений. Предложен алгоритм построения программы исследований, согласно которому пары скважин отбираются в программу на основе ответа классификатора — дерева решений. Поставлена задача повышения точности и контроля переобученности дерева решений. Исследованы причины возникновения переобучения классификатора, одной из которых является смещенность существующих критериев ветвления. Предложена модификация алгоритма построения дерева, согласно которойвыбор атрибута в узле проводится на основе комбинаторных оценок переобучения пороговых решающих правил, описанных в первой главе. Проведены вычислительные эксперименты на промысловых данных, которые показали статистически значимое повышение обобщающей способности дерева решений.

В четвертой главе решена задача разработки алгоритма для быстрого вычисления приближенных оценок обобщающей способности путем построения суррогатной модели. Актуальность задачи обоснована тем, чторазработанный полиномиальный алгоритмограничен быстрым увеличением вычислительной нагрузки с ростом мощностимножества классифицируемых объектов. Решение данной задачи основано на построении обучающей выборки для модели, которая состоит из пар «объект-ответ», где каждым объектом является семейство одномерных пороговых решающих правил, ответом является достигаемая верхняя оценка обобщающей способности семейства.

На основе имеющихся исследований оценок обобщающей способности, проведенных в рамках комбинаторной теории переобучения, был сформирован перечень признаков, которые описывают объекты выборки. Рассмотрены разные структурные модели, где наилучшей по результатам тестирования выбрана модель нейронной сети с MAPE=2.8% (среднее относительное отклонение приближенных оценок от их целевых значений).

Проведен анализ значимости признаков в модели, который подтвердил ранее полученные выводы, что при построении оценок необходимо учитывать внутреннюю структуру семейства: расслоение по числу ошибок и взаимосвязь между классификаторами (связность). Выделены признаки, которые можно использовать при построении оценок обобщающей способности для других семейств. Показано, что использование модели позволяет сократить время вычисления комбинаторных оценок зависимости

от мощности L генеральной совокупности на несколько порядков с $O(L^5)$ до $O(L^2)$ по сравнению с алгоритмом, построенным в первой главе.

В итоге сделан вывод о практической значимости разработанного подхода в задачах отбора признаков при построении деревьев решений, нейронных сетей и в алгоритмах бустинга для контроля переобучения.

Достоверность и обоснованность результатов. Достоверность результатов диссертации подтверждена строгими математическими доказательствами, экспериментальной проверкой разработанных методов в реальной задаче классификации скважин для планирования трассерных исследований нефтяных и газовых месторождений. Методика экспериментов изложена подробно, что обеспечивает возможность их воспроизведения. Основные положения, выносимые на защиту, опубликованы в рецензируемых научных изданиях, в том числе рекомендованных ВАК. Получен патент на изобретение и акт внедрения основных результатов. Основные результаты диссертации докладывались на восьми конференциях в 2013-23 гг.

Теоретическая и практическая значимость. В диссертации доказаны теоремы о вычислении достигаемых верхних оценок обобщающей способности одномерных пороговых решающих правил.

Теоретическая значимость результатов диссертационной работы заключается в том, что в рамках комбинаторного подхода до настоящего момента не удавалось получить достигаемые верхние оценки обобщающей способности для данного семейства в общем случае. Оценки были известны только для частного случая, где значения признака на классифицируемых объектах были попарно различны.

Практическая значимость результатов обосновывается тем, что разработанные оценки обобщающей способности применимы в задачах отбора признаков, в частности, при построении элементарных конъюнкций в логических алгоритмах классификации, решающих деревьев, решающих списков и при построении линейных классификаторов методом покоординатной оптимизации. Предложенный в работе способ построения программы трассерных исследований применим для повышения информативности исследований нефтегазовых месторождений.

Замечания. Имеются следующие замечания по тексту диссертации:

1. В гл. 1.4 на стр. 24-25 дается определение строгого порядка на парах классификаторов как бинарное отношение. Чуть выше на них же определяется условие их неразличимости. Здесь имеется формальное противоречие с определениями, т.к. в этом случае должен быть линейный квазипорядок;
2. В гл. 1.4 на стр. 24 в лемме 1.1 введены свойства 1 и 2 финитного отношения порядка, где, в соответствии с введенным отношением порядка $a \succ_X a'$ на стр. 25, утверждается, что предшествующий классификатор лучше второго на выборке X. Тогда, согласно данному определению, в

лемме 1.1 свойства 1 и 2 противоречивы, т.к. в правых частях должны быть одинаковые знаки $\leq$, как при сравнении числа ошибок, так и при сравнении переобученности двух классификаторов a_p и a_i. Тем не менее, далее по тексту диссертации становится понятным, что число ошибок (свойство 1) и переобученность (свойство 2) принципиально применяются в разных алгоритмах. В первом случае требуется минимизировать число ошибок для минимизации переобученности, а во втором случае – максимизировать переобученность для определения верхней оценки. Поэтому для устранения противоречия следует указать, что свойство 2 определено для обратного порядка на классификаторах;

3. В гл. 1.4 на стр. 25 в лемме 1.2 при определении свойства 2 финитного отношения порядка для полноты не хватает случая $i = p$ для нулевого запаса ошибок классификатора.

Опечатки. Следует отметить, что текст диссертации достаточно хорошо вычитан, в нем мало опечаток, что в современных условиях достаточно редко встречается:

1. В гл 1.1 на стр.16 при определении конечного множества **A**, элементы которого называются классификаторами, не указан его размер (в отличие от множества **X**);
2. Следует также обратить внимание, что обозначение […] занято на стр.16 под теоретико-множественное обозначение и тут же далее на стр. 17 оно же применяется для логических констант;
3. На стр. 17 используется необъявленное обозначение, где величина C_L^ℓ должна обозначать число сочетаний из L по l как мощность множества $[\boldsymbol{X}]^\ell$;
4. На этой же стр. 17 при определении функционала EOF в выражении справа от равенства следует заключить в скобки под знаком суммирования разность функций ν *ν*. Иначе, с учетом приоритета операции суммирования, формула теряет смысл, т.к. аргументы функции $\nu(\mu X, X), X \subset \boldsymbol{X}$ *ν*не определены. Семантически тут все понятно, но формальное выражение неверно;
5. Следует обратить внимание на выражение на стр. 47, гл. 2.3 для рademахеровских случайных величин σ_i – данное выражение не позволяет различать указанные ситуации;
6. На стр. 60 в формуле (3.4) вместо переменной $w_i(t)$ используется неизвестная переменная $w_{ij}(t)$;
7. На стр. 100 в ссылке 16 опечатка «закчки».

Принципиальных и других замечаний по содержанию работы, стилю и изложению материала нет. Все указанные замечания не влияют на общую положительную оценку диссертационной работы.

Заключение. Диссертационная работа Ш. Х. Ишкиной содержит решение актуальной задачи вычисления комбинаторных оценок переобучения одномерных пороговых решающих правил, связанной с разработкой новых подходов к оценке обобщающей способности решающих правил в машинном обучении, что соответствует паспорту специальности 1.2.1.

Основные результаты диссертации обладают научной новизной, имеют важное теоретическое и практическое значение. Они полностью опубликованы в 16 работах, в том числе в 8 статьях из списка изданий, рекомендованных ВАК, докладывались на четырех международных и четырех всероссийских конференциях. Автореферат адекватно отражает содержание диссертации.

Представленная диссертация является научной квалификационной работой, основные результаты которой следует рассматривать как решение актуальной задачи в области искусственного интеллекта и машинного обучения.

Диссертационная работа удовлетворяет всем требованиям, предъявляемым ВАК к диссертациям на соискание ученой степени кандидата физико-математических наук, а ее автор, Шаура Хабировна Ишкина, заслуживает присуждения ей ученой степени кандидата физико-математических наук по специальности 1.2.1 – «Искусственный интеллект и машинное обучение».

Официальный оппонент,
профессор, д.ф.-м.н., профессор С.Д. Двоенко

27.02.26

Адрес: 300012, Тула, пр. Ленина, 92, ТулГУ.
Тел.: 8 4872 25 79 40

Подпись Двоенко С.Д. заверяю
Ученый секретарь [illegible]