О комбинаторной теории качества обучения по прецедентам

К. В. Воронцов

24 января 2005

Вопрос о качестве алгоритмов, синтезированных по конечным выборкам прецедентов, является фундаментальной проблемой *теории обучаемых систем* (machine learning theory).

В общем случае задача обучения по прецедентам заключается в том, чтобы по заданной выборке пар «объект–ответ» восстановить функциональную зависимость между объектами и ответами, то есть построить алгоритм, способный выдавать адекватные ответы на предъявляемые объекты. Когда множество допустимых ответов конечно, говорят о задачах *классификации* или *распознавания образов*. Когда множество допустимых ответов бесконечно, например, является множеством действительных чисел или векторов, говорят о задачах *восстановления регрессии*. Когда объекты соответствуют моментам времени, а ответы характеризуют будущее поведение процесса или явления, говорят о задачах *прогнозирования*.

Значительный опыт решения прикладных задач распознавания, восстановления регрессии и прогнозирования был накоплен уже к середине 60-х годов. Большую популярность приобрёл подход, основанный на построении модели восстанавливаемой зависимости в виде параметрического семейства алгоритмов. С помощью численной оптимизации подбираются такие значения параметров, при которых алгоритм допускает наименьшее число ошибок на заданной обучающей выборке прецедентов. Проще говоря, осуществляется подгонка модели под выборку. Этот метод получил название *минимизации эмпирического риска*.

На практике исследователи столкнулись с проблемой *переобучения*. Чем больше у алгоритма свободных параметров, тем меньшего числа ошибок на обучении можно добиться путём оптимизации. Однако по мере нарастания сложности модели «оптимальные» алгоритмы начинают слишком хорошо подстраиваться под конкретные данные, улавливая не только черты восстанавливаемой зависимости, но и ошибки измерения обучающей выборки, и погрешность самой модели. В результате ухудшается качество работы алгоритма вне обучающей выборки, или, как говорят, его *способность к обобщению* (generalization ability).

Из этого наблюдения был сделан вывод, что для всякой задачи существует оптимальная сложность модели, при которой достигается наилучшее качество обобщения. Первое формальное обоснование этого практического опыта было дано в ста-

тистической теории восстановления зависимостей, разработанной В. Н. Вапником и А. Я. Червоненкисом в конце 60-х — начале 70-х [1, 2]. Эта теория получила широкую мировую известность и признание в середине 80-х. В настоящее время она активно развивается и используется для обоснования алгоритмов машинного обучения.

Основным результатом статистической теории являются количественные оценки, связывающие обобщающую способность алгоритмов с длиной обучающей выборки и сложностью семейства алгоритмов.

Основной проблемой статистической теории является чудовищная завышенность оценок. Непосредственный расчёт показывает, что для надёжного обучения необходимо иметь порядка 10^6–10^8 объектов. Это существенно превышает объёмы выборок, с которыми обычно приходится сталкиваться на практике. Тем не менее, прикладные задачи решаются, и вполне успешно. Наиболее интересные случаи — малых выборок и сложных семейств алгоритмов — находятся за границами применимости теории. По сути дела, теория дает лишь качественное обоснование некоторых эвристических принципов построения обучаемых алгоритмов.

Основной причиной завышенности статистических оценок является их чрезмерная общность. Они не учитывают существенных особенностей метода обучения, восстанавливаемой зависимости и распределения объектов в пространстве. Иными словами, теория пессимистично настроена на «худший случай», который почти невозможно встретить на практике. Почему — вопрос философский. Одни философы говорят «Бог изощрён, но не злонамерен». Другие объясняют это тем, что человеческая деятельность настолько проста и поверхностна в этом мире, что мы просто не в состоянии поставить перед собой по-настоящему сложную задачу.

Основная цель дальнейших исследований — довести точность оценок до уровня практической применимости. В идеале они должны предсказывать частоту ошибок построенного алгоритма примерно с той же точностью, с которой закон больших чисел предсказывает частоту выпадения орла или решки. Только тогда на основе этих оценок можно будет целенаправленно конструировать алгоритмы высокого качества.

Комбинаторный подход [3, 4, 7, 5, 6] возник как попытка более аккуратного построения статистической теории обучения, начиная с исходных её постулатов. Толчком для этого послужило следующее наблюдение.

В любой теории имеется, по выражению Пойа, «скрытая движущая пружина», некий момент доказательства, в который «всё и происходит». В статистической теории тоже есть такая теорема [2, стр. 221], и в ней хорошо видны все моменты, в которые происходит потеря точности оценки. Удивительно, что доказательство является по сути комбинаторным, а не вероятностным. Вероятностная мера используется лишь в самом начале доказательства, когда в оцениваемый функционал искусственно вводится суммирование по всевозможным разбиениям выборки. Чтобы этот трюк проходил, выборка должна быть случайной, независимой и одинаково распределённой. То есть вероятностные предположения нужны только для того, чтобы подменить функционал, и на существо дела не влияют. Так какой же функционал на самом деле оценивается в основной теореме Вапника-Червоненкиса? Если разобраться,

то оказывается, что это некоторый вариант хорошо известного функционала скользящего контроля. Точнее, это доля разбиений выборки на две части, обучающую и контрольную, при которых частота ошибок на контроле существенно превосходит частоту ошибок на обучении. Кстати, аналогичную теорему удалось доказать и для более традиционного варианта скользящего контроля — функционала средней (по всем разбиениям выборки) частоты ошибок на контроле [7].

Интересно, что теоремы вапниковского типа оказываются справедливы для функционалов качества, не содержащих в своём определении понятия вероятности. Вместо вероятности ошибки они оценивают частоту ошибок на произвольной контрольной выборке. По смыслу эта величина ничуть не хуже характеризует обобщающую способность. При этом комбинаторные оценки могут быть преобразованы обратно в вероятностные «одним действием» — взятием матожидания от левой и правой частей неравенства.

Снимается ли при этом требование случайности и независимости выборки? Случайности — да, независимости — отчасти. В теории вероятностей независимость означает инвариантность вероятностной меры относительно всевозможных перестановок выборки. В комбинаторном подходе ту же роль в доказательствах играет свойство инвариантности функционала качества относительно всевозможных перестановок выборки. Назовём это свойство симметричностью функционала. Требование симметричности можно считать слабой формой гипотезы независимости, при которой ограничение переносится с исходных данных на функционал качества. Это действительно ослабление ограничений, поскольку исходные данные приходят извне, а функционал мы вправе выбирать сами.

Основная гипотеза вероятностного подхода звучит так: «мы верим, что выборка является случайной, независимой, одинаково распределённой». Неявно предполагается ещё больше: «мы верим, что множество всех возможных объектов образует вероятностное пространство, в котором существует сигма-аддитивная мера, и что все используемые случайные величины являются измеримыми функциями». На практике эти предположения, как правило, не проверяются. Их вообще трудно проверить эмпирически.

Основная гипотеза комбинаторного подхода звучит по-другому: «мы верим, что скользящий контроль адекватно оценивает качество обучения по прецедентам». Эта гипотеза подтверждается эмпирическим опытом нескольких поколений исследователей, работающих в области статистики и машинного обучения. Именно скользящий контроль (в различных модификациях, что не так уж важно) принято использовать для эмпирического сравнения качества алгоритмов. Альтернативных методик просто нет.

Итак, комбинаторный подход подразумевает отказ от вероятностной трактовки задачи обучения по прецедентам.

Вообще, современная теория вероятностей возникла из стремления объединить в рамках единого формализма частотное понятие вероятности, берущее начало от азартных игр, и континуальное, идущее от геометрических задач типа задачи

Бюффона. В аксиоматике Колмогорова континуальное понятие берётся за основу как более общее. Ради этой общности в теорию вероятностей привносятся технические предположения из теории меры, имеющие довольно слабые эмпирические обоснования (гипотезы сигма-аддитивности и измеримости). Однако далеко не во всех задачах, связанных со случайностью, определение вероятности как меры действительно необходимо. Для описания существенно дискретных явлений, более похожих на азартные игры, чем на попадание иглы в паркетную щель, применение теории меры представляется избыточным.

На практике обучаемые алгоритмы имеют дело только с конечными выборками, будь то обучающие или контрольные совокупности объектов. Следовательно, по своей природе задача обучения является скорее дискретной, чем континуальной. Для адекватного описания и исследования таких задач вполне достаточно «частотного раздела» теории вероятностей, которую можно рассматривать как раздел комбинаторики.

Комбинаторная перестройка аксиоматики приводит к пересмотру многих положений статистической теории обучения. Основное изменение заключается в следующем. В каждой конкретной задаче восстанавливаемая зависимость и метод обучения фиксированы, обучающая выборка конечна. Поэтому лишь ограниченная (локальная) часть семейства алгоритмов может быть получена в результате обучения. Остальные алгоритмы остаются незадействованными. Этот эффект назван «локализацией» семейства алгоритмов. Качество обучения зависит от не столько от ёмкости всего семейства, сколько от локализующей способности конкретного метода обучения. Таким образом, в комбинаторной теории появляется принципиальная возможность снять искусственный запрет на использование сложных алгоритмов.

Ещё одно преимущество комбинаторной техники в том, что она позволяет более аккуратно проследить цепочку неравенств от функционала к финальной оценке. Некоторые переходы в ней просто отсутствуют, например, переход от разности частоты ошибок и их вероятности к разности частот ошибок в двух подвыборках. Даже само введение функционала равномерной сходимости уже является оценкой, тогда как в комбинаторной теории требование равномерной сходимости даже не возникает. Для устранения завышенности оценок необходимо разобраться, сколько мы теряем в точности на каждом знаке $\leqslant$ в цепочке неравенств. В современном изложении статистической теории обучения такие переходы совершаются бесконтрольно [8].

Комбинаторная техника позволяет устранить не все источники завышенности вапниковских оценок. Сам переход от качества к сложности в процессе доказательства основной теоремы неизбежно приводит к существенному ухудшению оценок. Отсюда гипотеза: любые сложностные оценки принципиально завышены. И вывод: для получения оценок, непосредственно применимых на практике, необходимо учитывать не только сложность семейства алгоритмов, но и более тонкие характеристики метода обучения, а также привлекать всю доступную априорную информацию о выборке и восстанавливаемой зависимости.

Оценки вапниковского типа зависят только от ёмкости семейства, частоты оши-

бок на обучающей выборке и её длины. Три скалярные характеристики едва ли могут содержать достаточно информации о таком сложном процессе, как обучение по прецедентам. Для более точного оценивания просто не хватает информации!

Оказывается, дополнительная информация может быть получена как «побочный продукт» процесса обучения. Вместо скалярной частоты ошибок на обучающей выборки в ходе обучения вычисляется некоторая векторная характеристика, называемая *профилем*. К сожалению, универсального определения профиля пока не найдено. Как сам профиль, так и оценки обобщающей способности, выражающиеся через профиль, существенно зависят от метода обучения. Сейчас можно указать 6 принципиально различных типов методов, для которых возникает понятие профиля.

Профиль компактности возникает в метрических алгоритмах классификации, таких как алгоритм ближайших соседей. Профиль показывает, насколько часто близкие объекты попадают в один класс.

Профиль разделимости возникает в дискриминантных алгоритмах классификации, основанных на явном построении поверхностей, разделяющих классы. Профиль показывает, как объекты обучения распределены по расстояниям до разделяющей поверхности.

Профиль монотонности возникает в алгоритмах, удовлетворяющих дополнительному ограничению монотонности. Такие алгоритмы используются, в частности, как корректирующие операции при построении алгоритмических композиций. Профиль показывает, как меняется плотность отношения порядка по мере приближения к границе классов.

Профиль устойчивости возникает для широкого класса методов, которые при малом изменение состава обучающей выборки строят почти одинаковые алгоритмы. Профиль показывает, какова доля объектов, на которых изменяется ответ алгоритма, при замене заданного числа обучающих объектов.

Профиль информативности возникает в логических алгоритмах классификации, основанных на поиске логических закономерностей. Профиль показывает, сколько предикатов с заданной информативностью было просмотрено в процессе поиска закономерностей.

Профиль альтернатив возникает в задачах выбора наилучшей модели алгоритмов из небольшого конечного множества альтернативных моделей. Профиль показывает распределение частоты ошибок моделей на контрольных данных, по которым производится выбор моделей.

Такой подход позволяет строить обучаемые алгоритмы по принципу явной оптимизации профиля и управлять качеством алгоритма в процессе его построения. По окончании процесса выдается не только сам алгоритм, но и оценка его обобщающей способности. В отличие от статистической теории, эта оценка уже не может быть получена заранее по известной формуле, поскольку она вычисляется в процессе обучения и влияет на этот процесс.

Список литературы

[1] *Вапник В. Н., Червоненкис А. Я.* Теория распознавания образов. — М.: Наука, 1974.

[2] *Вапник В. Н.* Восстановление зависимостей по эмпирическим данным. — М.: Наука, 1979.

[3] *Воронцов К. В.* Качество восстановления зависимостей по эмпирическим данным // Математические методы распознавания образов: 7-ая Всерос. конф. Тезисы докл. — Пущино, 1995. — С. 24–26.

[4] *Воронцов К. В.* О комбинаторном подходе к оценке качества обучения алгоритмов // Математические методы распознавания образов: 11-ая Всерос. конф. Тезисы докл. — Пущино, 2003. — С. 47–49.

[5] *Воронцов К. В.* Комбинаторные обоснования обучаемых алгоритмов // *ЖВМиМФ.* — 2004. — Т. 44, № 11. — С. 2099–2112.
`http://www.ccas.ru/frc/papers/voron04jvm.pdf`.

[6] *Воронцов К. В.* Комбинаторные оценки качества обучения по прецедентам // *Докл. РАН.* — 2004. — Т. 394, № 2. — С. 175–178.
`http://www.ccas.ru/frc/papers/voron04qualdan.pdf`.

[7] *Воронцов К. В.* Комбинаторный подход к оценке качества обучаемых алгоритмов // *Математические вопросы кибернетики.* — 2004. — № 13. — С. 5–36.
`http://www.ccas.ru/frc/papers/voron04mpc.pdf`.

[8] *Bousquet O., Boucheron S., Lugosi G.* Theory of classification. — 1999.