Толстихин Илья: Краткое изложение заявки

Оценки обобщающей способности и локальные меры сложности в теории машинного обучения

В теории *машинного обучения* одной из важных задач является строгое математическое обоснование и анализ поведения *обучаемых алгоритмов* — закономерностей, выявленных на наблюдаемых *эмпирических* данных. Этим вопросам посвящен один из наиболее математизированных разделов машинного обучения — *теория статистического обучения* (statistical learning theory), основы которой были заложены в работах советских ученых В. Н. Вапника и А. Я. Червоненкиса 1970-х годов. Несмотря на то, что это передовое направление теории обучения машин в настоящее время бурно развивается на Западе, в России ему уделяется незаслуженно мало внимания.

В последние годы в *теории эмпирических процессов* было получено множество результатов, позволивших улучшить классические оценки обобщающей способности классификаторов. Неравенство Талаграна для эмпирических процессов, опубликованное в 96-м году, дало толчок ряду новых подходов в этой области.

Классические результаты в рамках вероятностной постановки задачи обучения сводили оценку среднего риска (математического ожидания) функции f, выбираемой методом обучения из класса $\mathcal{F}$, к оценке величины равномерного по классу $\mathcal{F}$ отклонения *эмпирического риска* на случайной выборке $\{X_i\}_{i=1}^n$ от математического ожидания: $\sup_{f\in\mathcal{F}} |\frac{1}{n}\sum_{i=1}^n f(X_i) - \mathsf{E}f|$. Для применяемых на практике методов обучения эта оценка может быть сильно завышенной в результате взятия супремума по всему классу $\mathcal{F}$, в то время как достаточно брать его по небольшому подмножеству класса $\mathcal{F}$, куда входят функции с малым средним риском.

За последние годы такими учеными как Koltchinskii, Panchenko и Bartlett был развит подход, основанный на *локализации* — замене всего класса под знаком супремума на некоторые его подмножества. Такая замена в ряде случаев ведет к существенным улучшениям оценок обобщающей способности.

Комбинаторная теория надежности обобщения по прецедентам, предложенная К. В. Воронцовым и активно развивающаяся в настоящий момент, также позволяет учитывать локализацию для случая бинарных функций потерь, давая возможность получать даже точные (не завышенные, не ассимптотические) оценки для некоторых классов функций $\mathcal{F}$.

Автором в рамках комбинаторного подхода к оценкам обобщающей способности был получен ряд точных оценок для нескольких модельных семейств классификаторов: хэммингова шара, одного слоя хэммингова шара и его нижних слоев. Также велся анализ некоторых поздних результатов из теории эмпирических процессов в терминах комбинаторного подхода, в ходе которого был доказан аналог *симметризации* — одного из широко используемых результатов в теории статистического обучения — для *рademахеровской сложности* в рамках комбинаторного подхода.

Целью данного проекта является формирование более глубокого понимания природы процесса обучения. Это позволит развить новые методы обучения с контролируемой обобщающей способностью, что приведет к получению более хороших классификаторов. Автор собирается посвятить свою работу следующим открытым вопросам:

- Комбинаторный подход к оценке обобщающей способности продолжает активно развивается в России. Однако, на данный момент не до конца понятно, как он соотносится со свежими результатами теории эмпирических процессов и теории статистического обучения, полученными на Западе. Помимо этого открытыми вопросами остаются следующие:
  1) возможность перехода в рамках подхода от бинарных функций потерь к вещественным;
  2) переход от ненаблюдаемых оценок, зависящих от скрытых параметров семейства $\mathcal{F}$, к наблюдаемым, которые можно вычислить по обучающей выборке $\{X_i\}_{i=1}^n$.

  По мнению автора, перспективный путь к изучению поставленных вопросов лежит в теории эмпирических процессов и теории *концентрации меры*.

- Последние результаты, связанные с локализацией, дают достаточно точные *ненаблюдаемые* оценки обобщающей способности для классов функций, зависящие от неизвестного вероятностного распределения P (distribution-dependant). Оказывается, в общем случае невозможно получить сравнимые по точности *наблюдаемые* оценки, вычислимые с помощью обучающей выборки (data-dependant). Открытым вопросом является изучение ограничений на класс функций и вероятностное распределение, позволяющих преодолевать эту сложность.