АНАЛИЗ СХОДСТВА АЛГОРИТМОВ КЛАССИФИКАЦИИ В ОЦЕНКАХ ОБОБЩАЮЩЕЙ СПОСОБНОСТИ

Цюрмасто П. А., Воронцов К. В.
МФТИ, Вычислительный Центр РАН**, Россия*
e-mail: scrime69pete@mail.ru, voron@ccas.ru

Большинство из известных верхних оценок вероятности переобучения алгоритмов классификации имеют следующий вид [1,3,4]:

$$P\{\nu > \hat{\nu}_m + \varepsilon\} \le \Delta \cdot \Gamma(m, \varepsilon), \qquad (1)$$

где $\hat{\nu}_m$ — частота ошибок алгоритма, построенного по обучающей выборке длины m (эмпирический риск); ν — вероятность ошибки этого алгоритма, либо частота его ошибок на неизвестной контрольной выборке; $\Gamma(m, \varepsilon)$ — экспоненциально убывающий по m множитель, фактически, оценка скорости сходимости в законе больших чисел; Δ — показатель сложности используемого семейства алгоритмов (число алгоритмов, функция роста, shattering coefficient, и др).

Известно, что в оценках такого вида показатель Δ завышен, как правило, на много порядков. Причин завышенности несколько. Во-первых, в каждой задаче целевая зависимость фиксирована, поэтому реально используется не всё множество алгоритмов, а лишь его локальное подмножество. Оценки, зависящие от выборки данных, активно исследуются в последние годы [3,4], но и они пока остаются завышенными на несколько порядков.

Вторая причина в том, что в локальном подмножестве алгоритмов многие алгоритмы схожи.

Рассмотрим простой частный случай. Пусть имеются два алгоритма, каждому из них соответствует бинарный вектор ошибок на выборке длины L, и пусть d — хэммингово расстояние между этими векторами. Выборка всеми возможными C_L^m способами равновероятно разбивается на две части: обучающую длины m и контрольную длины $k = L - m$. В данной работе найдена вероятность P того, что алгоритм с меньшей частотой ошибок $\hat{\nu}_m$ на обучающей части имеет частоту ошибок на контрольной части ν, превышающую $\hat{\nu}_m + \varepsilon$. Эта вероятность также имеет вид (1). Исследована зависимость показателя Δ от различности алгоритмов d. Показано, что $1 \le \Delta \le 2$, причём с уменьшением d до нуля Δ уменьшается до единицы. Иными словами, сложность семейства, состоящего из двух алгоритмов, близка к 1, а не к 2, если эти алгоритмы схожи. Оценка вероятности P получена

чисто комбинаторными методами, является точной, и имеет довольно сложное выражение.

Попытка рассмотреть тем же способом случай хотя бы трёх алгоритмов наталкивается на значительные технические трудности. Поэтому следующим был рассмотрен частный случай *цепочки алгоритмов*, в которой для любых двух соседних алгоритмов хеммингово расстояние между их векторами ошибок равно единице. Такие цепочки возникают в алгоритмах классификации с непрерывной по параметрам дискриминантной функцией. Это очень широкий класс алгоритмов, поэтому данный частный случай может служить основой для дальнейших обобщений. Заметим, что попытки учесть сходство алгоритмов при оценивании ёмкости семейства предпринимались и ранее [2,5], но остались практически незамеченными.

В данной работе получена комбинаторная верхняя оценка вероятности переобучения для случая, когда алгоритм выбирается из цепочки по критерию минимума эмпирического риска. Оценка не имеет простого аналитического выражения и вычисляется с помощью рекуррентного алгоритма. Предварительные имитационные эксперименты со скользящим контролем показали, что предложенная оценка почти не завышена. Сопоставление вычисляемой оценки с оценкой (1) показывает, что зависимость Δ от длины цепочки растёт существенно медленнее линейной. Отсюда следует вывод, что характеристики сложности Δ, такие как функция роста или ёмкость, основанные на простом подсчёте числа алгоритмов без учёта степени их различности, ведут к сильно завышенным оценкам обобщающей способности (вероятности переобучения).

Работа выполнена при поддержке РФФИ, проект 08-07-00422.

Литература

1. *Воронцов К. В. Комбинаторный подход к оценке качества обучаемых алгоритмов // Математ. вопросы кибернетики / Под ред. О. Б. Лупанова. — М.: Физматлит, 2004. — Т. 13. — С. 5–36.*
2. *Bax E. T. Similar classifiers and VC error bounds: Tech. Rep. CalTech-CS-TR97-14: 6 1997.*
3. *Boucheron S., Bousquet O., Lugosi G. Theory of classification: A survey of some recent advances // ESAIM: Probability and Statistics. — 2005. — no. 9. — Pp. 323–375.*
4. *Langford J. Quantitatively tight sample complexity bounds. — 2002. — Carnegie Mellon Thesis.*
5. *Sill. J. Monotonicity and connectedness in learning systems. —1998. — Caltech Thesis.*