Слабая вероятностная аксиоматика и надёжность эмпирических предсказаний

Воронцов К. В.
voron@ccas.ru
Москва, Вычислительный центр РАН

Задача эмпирического предсказания является одной из центральных в прикладной статистике и машинном обучении: получив выборку данных, необходимо предсказать определённые свойств аналогичных данных, которые станут известны позже, и оценить точность предсказания. В сообщении предлагается новая формализация постановки задачи, не требующая привлечения классической вероятностной аксиоматики.

Пусть задано множество объектов $\mathbb{X}$ и выборка $X^L = (x_1, \ldots, x_L) \subseteq \mathbb{X}$ длины L. Рассмотрим множество всех её разбиений на две подвыборки длины ℓ и k соответственно: $X^L = X^\ell_n \cup X^k_n,\ \ell + k = L$, где нижний индекс $n = 1, \ldots, N$ пробегает все $N = C^k_L$ разбиений.

Пусть задано множество R и функция $T\colon \mathbb{X}^* \times \mathbb{X}^* \to R$, где $\mathbb{X}^*$ — множество всех конечных выборок из $\mathbb{X}$.

Рассмотрим эксперимент, в котором с равной вероятностью реализуется одно из разбиений n, после чего наблюдателю сообщается выборка X^ℓ_n. Не зная *скрытой выборки* X^k_n, наблюдатель должен построить функцию $\hat{T}\colon \mathbb{X}^* \to R$, значение которой на *наблюдаемой выборке* $\hat{T}_n = \hat{T}(X^\ell_n)$ предсказывало бы значение $T_n = T(X^k_n, X^\ell_n)$, существенно зависящее от скрытой выборки X^k_n. Требуется также оценить надёжность предсказания, т. е. указать *оценочную функцию* $\eta(\varepsilon)$ такую, что

$$\mathsf{P}_n\big\{d(\hat{T}_n, T_n) > \varepsilon\big\} \leqslant \eta(\varepsilon), \tag{1}$$

где $d\colon R \times R \to \mathbb{R}$ — заданная функция, характеризующая величину отклонения $d(\hat{r}, r)$ предсказанного значения $\hat{r} \in R$ от неизвестного истинного значения $r \in R$. Параметр ε называется *точностью*, а величина $(1 - \eta(\varepsilon))$ — *надёжностью* предсказания. Если в (1) достигается равенство, то $\eta(\varepsilon)$ называется *точной оценкой*. Оценка $\eta(\varepsilon)$ может зависеть от ℓ и k, а также от вида функций T и $\hat{T}$. Если (1) выполняется при достаточно малых ε и η, то говорят, что в окрестности предсказываемого значения имеет место *концентрация вероятности* [5].

Заметим, что данная постановка задачи не опирается на классическую аксиоматику теории вероятностей. Здесь понятие вероятности является лишь синонимом доли разбиений: $\mathsf{P}_n\{\varphi(n)\} = \frac{1}{N}\sum_{n=1}^{N}\varphi(n)$ для произвольного предиката $\varphi\colon \{1, \ldots, N\} \to \{0, 1\}$, заданного на множестве разбиений выборки X^L. Тем не менее, мы предпочитаем пользоваться привычным термином *вероятность* и говорить, что задача эмпирического предсказания поставлена в *слабой вероятностной аксиоматике*.

Слабая аксиоматика ориентирована на задачи анализа данных, в которых все выборки конечные и все величины наблюдаемые, т. е. являются функциями конечных выборок. В классической колмогоровской аксиоматике вероятность события, функция распределения и матожидание случайной величины являются величинами ненаблюдаемыми. В задачах анализа данных слабая аксиоматика имеет ряд преимуществ.

1. Упрощается понятийный аппарат. Нет необходимости использовать теорию меры, предельный переход к бесконечной выборке, различные типы сходимости, и т. д. Однако это не мешает сформулировать и доказать аналоги многих фундаментальных утверждений теории вероятностей и математической статистики: закон больших чисел, сходимость эмпирических распределений (критерий Смирнова), ранговые критерии, оценки Вапника-Червоненкиса [6], и т. д.

2. Сильная (колмогоровская) аксиоматика требует, чтобы на множестве объектов $\mathbb{X}$ существовала σ-аддитивная алгебра событий, объекты X^L выбирались случайно из фиксированной генеральной совокупности, и все рассматриваемые функции выборок были измеримы. Требования случайности, независимости и одинаковой распределённости могут быть проверены с помощью статистических тестов. Однако гипотезы σ-аддитивности и измеримости эмпирической проверке не поддаются [1]. Слабая аксиоматика обходится без этих гипотез. Фактически, в ней остаётся только гипотеза равновероятности разбиений, эквивалентная предположению о независимости выборки X^L. Об объектах вне выборки X^L вообще не делается никаких предположений.

3. Из оценки «слабого» функционала $\mathsf{P}_n\{\varphi(X_n^\ell, X_n^k)\} \leqslant \eta$ всегда можно получить оценку «сильного» функционала, взяв матожидание по выборке X^L от обеих частей неравенства:

$$\mathsf{E}_{X^L}\,\mathsf{P}_n\{\varphi(X_n^\ell, X_n^k)\} = \mathsf{P}_{X^L}\{\varphi(X^\ell, X^k)\} \leqslant \mathsf{E}_{X^L}\eta.$$

Если η не зависит от выборки, то оценка переносится непосредственно. Для оценок типа Вапника-Червоненкиса это было проделано в [3].

4. С другой стороны, «слабые» функционалы легко поддаются эмпирическому измерению. Для этого суммирование по всем разбиениям заменяется суммированием по некоторому подмножеству разбиений (в методе Монте-Карло — по случайному). Таким образом, слабая аксиоматика является единой отправной точкой как для теоретико-вероятностного, так и для экспериментального анализа надёжности эмпирических предсказаний. В теории машинного обучения становится предельно понятной связь между теоретическими оценками обобщающей способности и практическими методиками, основанными на скользящем контроле.

5. В современной вычислительной теории обучения [5] для получения верхних оценок надёжности используется математический аппарат функционального анализа и оценки концентрации вероятностной меры. Это мощная и красивая математическая теория, но, к сожалению, в ходе вывода оценок их точность теряется практически бесконтрольно на многочисленных промежуточных шагах. Проблема в том, что «асимптотичность» заложена как в самом понятии вероятности, так и в стремлении получить оценку в виде «изящной» формулы, даже ценой значительной потери её точности. Слабая аксиоматика во многих случаях приводит к точным, не асимптотическим, оценкам. Иногда это довольно сложные комбинаторные выражения, требующие значительных объёмов вычислений. Однако во многих случаях находятся эффективные алгоритмические решения вычислительных проблем.

Вышесказанное позволяет выдвинуть смелую гипотезу: *для исследования задач эмпирического предсказания достаточно слабой вероятностной аксиоматики.* Пока вопрос о границах её применимости остаётся открытым. Приведём два высказывания, подтолкнувших автора к проведению данного исследования.

А. Н. Колмогоров [4]: «... представляется важной задача освобождения всюду, где это возможно, от излишних вероятностных допущений. На независимой ценности чисто комбинаторного подхода к теории информации я неоднократно настаивал в своих лекциях».

Ю. К. Беляев [2]: «... возникло глубокое убеждение, что в теории выборочных методов можно получить содержательные аналоги большинства основных утверждений теории вероятностей и математической статистики, которые к настоящему времени найдены в предположении взаимной независимости результатов измерений».

Работа выполнена при поддержке РФФИ, проект №05-01-00877, и программы ОМН РАН «Алгебраические и комбинаторные методы математической кибернетики».

Литература

[1] *Алимов Ю. И.* Альтернатива методу математической статистики. — Знание, 1980.

[2] *Беляев Ю. К.* Вероятностные методы выборочного контроля. — М.: Наука, 1975.

[3] *Воронцов К. В.* Комбинаторный подход к оценке качества обучаемых алгоритмов // Математические вопросы кибернетики / Под ред. О. Б. Лупанова. — М.: Физматлит, 2004. — Т. 13. — С. 5–36.

[4] *Колмогоров А. Н.* Теория информации и теория алгоритмов / Под ред. Ю. В. Прохорова. — М.: Наука, 1987.

[5] *Lugosi G.* On concentration-of-measure inequalities. // Machine Learning Summer School, Australian National University, Canberra. — 2003.

[6] *Vapnik V.* Statistical Learning Theory. — Wiley, New York, 1998.