МАКСИМАЛЬНОЕ СМЕЩЕНИЕ ВЫБОРОЧНОЙ ОЦЕНКИ РИСКА РЕШАЮЩЕЙ ФУНКЦИИ ДЛЯ ДИСКРЕТНОГО ПРОСТРАНСТВА

В. М. Неделько[1]

Работа посвящена проблеме статистической устойчивости решающих функций, или оценивания риска. Для случая дискретной "независимой" переменной получена точная зависимость между средними риском и эмпирическим риском решения для "наихудшего" распределения. Полученный результат позволяет для рассмотренного случая непосредственно оценить степень завышенности оценок Вапника–Червоненкиса.
Предлагается эвристический подход к применению полученных результатов к оцениванию качества решения для реальных задач.

Введение

Из причин завышенности оценок Вапника–Червоненкиса наиболее часто выделяют следующие две: переоценка богатства (емкости) реально используемого класса решающих функций и ориентацию оценок на «наихудшее» распределение. В данной работе внимание сосредоточено на выявлении погрешности, вносимой техническим аппаратом, использованным при получении оценок. В этой части основным источником погрешности представляется заложенное в этой технике оценивание вероятности суммы событий через сумму их вероятностей.

В работе дается точное решение экстремальной задачи нахождения наибольшего (по всем распределениям) уклонения риска от эмпирического риска для случая дискретной переменной, что позволяет в рассмотренном частном случае точно вычислить погрешность оценок Вапника–Червоненкиса.

Пусть X – пространство значений "независимых переменных", а Y – целевое пространство прогнозируемых значений. Пусть C – множество вероятностных мер на $D = X \times Y$. Конкретную меру $c \in C$ будем обозначать $P_c[D]$.

Здесь и далее квадратные скобки будут использоваться для указания множества, на σ-алгебре подмножеств которого задается мера, а круглые скобки — для указания меры множества (вероятности события).

В общем случае необходимо дополнительно потребовать также существование $\forall x \in X$ условных мер $P_c[Y/x]$.

Под решающей функцией понимается отображение $f : X \to Y$.

Для определения качества решающей функции необходимо задать функцию потерь: $L : Y^2 \to [0, \infty)$.

Теперь под риском будем понимать средние потери:

$R(c, f) = \int L(y, f(x)) dP_c[D] = \int R_x(c, f)\, dP_c[X]$,

где $R_x(c, f) = \int L(y, f(x))\, dP_c[Y/x]$.

[1] Институт математики СО РАН, 630090, Россия, Новосибирск, пр-т Коптюга, 4, тел.: (3832)332892, nedelko@math.nsc.ru.

Решающая функция строится на основе случайной независимой выборки $v_c = \left\{ \left(x^i, y^i\right) \in D \,\middle|\, i \in \overline{1,N} \right\}$ из распределения $P_c[D]$.

В качестве выборочной оценки риска вводится понятие эмпирического риска: $\tilde{R}(v,f) = \frac{1}{N}\sum_{i=1}^{N} L\left(y^i, f\left(x^i\right)\right)$.

Для всех практически используемых алгоритмов построения решающих функций эмпирический риск является смещенной оценкой риска, причем всегда заниженной, поскольку данные алгоритмы минимизируют эмпирический риск.

Актуальной является задача оценки этого смещения.

Введем обозначения: $F(c,Q) = ER\left(c, f_{Q,v}\right)$, $\tilde{F}(c,Q) = E\tilde{R}\left(c, f_{Q,v}\right)$.

Здесь $Q : \{v\} \to \{f\}$ алгоритм построения решающей функции, а $f_{Q,v}$ – решающая функция, построенная по выборке *v* алгоритмом *Q*.

Математическое ожидание берется по всем выборкам объема *N*.

Введем функцию наибольшего смещения:

$$S_Q\left(\tilde{F}_0\right) = \hat{F}_Q\left(\tilde{F}_0\right) - \tilde{F}_0, \qquad (1)$$

где $\hat{F}_Q\left(\tilde{F}_0\right) = \sup_{c:\tilde{F}(c,Q)=\tilde{F}_0} F(c,Q)$.

Основной результат данной работы заключается в нахождении зависимости $S_Q\left(\tilde{F}_0\right)$ для случая дискретного *X* и *Q,* минимизирующего эмпирический риск в каждой точке *x*.

Дискретный случай

Пусть *X* дискретно, то есть $X = \{1,\ldots,n\}$.

Тогда $R(c,f) = \sum_{x=1}^{n} p_x R_x\left(\xi_x, f(x)\right)$, где $R_x\left(\xi_x, v_x\right) = P\left(y \neq f(x)/x\right)$ – условный риск в точке *x*, $p_x = P_c(x)$, ξ_x – краткое обозначение условной меры $P_c[Y/x]$, v_x – подмножество выборки для значения x.

Зададимся решающей функцией, минимизирующей эмпирический риск: $f_v^*(x) = \arg\min_{y \in Y} \tilde{L}(y, v_x)$, $\tilde{L}(y, v_x) \sum_{v_x} L\left(y, y^i\right)$.

Определим, от каких величин зависят искомые средние: $F(c,Q) = ER\left(c, f_v^*\right) = \sum_{x=1}^{n} F_x\left(p_x, \xi_x\right)$, где $F_x\left(p_x, \xi_x\right) = p_x ER_x\left(\xi_x, f_{v_x}^*(x)\right)$.

Аналогично,

$\tilde{F}(c,Q) = \frac{1}{N} E\tilde{L}\left(v, f_v^*\right) = \sum_{x=1}^{n} \tilde{F}_x\left(p_x, \xi_x\right)$ где $F_x\left(p_x, \xi_x\right) = E\tilde{L}_x\left(v_x, f_{v_x}^*(x)\right)$.

Введем функцию $\hat{F}_x\left(p_x, \tilde{F}_x^0\right) = \sup_{\xi_x : \tilde{F}_x(p_x,\xi_x) = \tilde{F}_x^0} F_x\left(p_x, \xi_x\right)$.

Теперь

$$\hat{F}_Q\left(\tilde{F}_0\right) = \max \sum_{x=1}^{n} \hat{F}_x\left(p_x, \tilde{F}_x^0\right), \qquad (2)$$

где максимум берется по всем p_x и $\widetilde{F}_x^0$, $x=\overline{1,n}$ при ограничениях: $p_x \ge 0$, $\widetilde{F}_x^0 \ge 0$, $\sum_{x=1}^{n} p_x = 1$, $\sum_{x=1}^{n} \widetilde{F}_x^0 = \widetilde{R}_0$.

Исходная экстремальная задача существенно упростилась, разбившись на два этапа: нахождение $\hat{F}_x\left(p_x, \widetilde{F}_x^0\right)$ и нахождение максимума функции на простой подобласти Евклидова пространства.

Функция $\hat{F}_x\left(p_x, \widetilde{F}_x^0\right)$, как показано ниже, может быть легко аппроксимирована путем численного моделирования.

Однако решать задачу (2) непосредственно численными методами проблематично, поскольку размерность пространства, по которому производится максимизация, равна $2n$ где n может быть достаточно велико (практический интерес представляют значения порядка десятков).

Решение экстремальной задачи

Перепишем задачу (2) в абстрактных обозначениях:

$$\sum_{x=1}^{n} \Phi\left(z^x\right) \to \max_{z^x}, \tag{3}$$

$$z^x = \left(z_1^x, ..., z_m^x\right),\ z_j^x \ge 0,\ \sum_{x=1}^{n} z_j^x = 1,\ j=\overline{1,m}.$$

Применительно к задаче (2), Φ соответствует $\hat{F}_x$, $m=2$, $z_1^x = p_x$, $z_2^x = \frac{\widetilde{F}_x^0}{\widetilde{F}_0}$.

Дискретизуем пространство значений вектора z^x. То есть предположим, что $z^x \in \left\{t^1, ..., t^l\right\}$.

Тогда задачу (3) можно переписать в эквивалентном виде:

$$\sum_{i=1}^{l} \Phi\left(t^i\right)\kappa^i \to \max_{\kappa^i}, \tag{4}$$

$$\kappa^i \ge 0,\ \sum_{i=1}^{l} t_j^i \kappa^i = 1,\ j=\overline{1,m},\ \sum_{i=1}^{l} \kappa^i = 1, \text{ где } \kappa^i \in \left\{ \frac{a}{n} \,\middle|\, a=\overline{0,n} \right\}.$$

Будем решать задачу без последнего ограничения дискретности κ^i.

Получаем задачу линейного программирования, решение которой находится в некоторой вершине области значений аргументов. Это означает, что только m+1 из κ^i могут быть ненулевыми.

Теперь легко показать, что вносимая снятием ограничения дискретности κ^i погрешность оказывается порядка $\frac{m}{n}$, и ей можно пренебречь.

Поскольку вывод о числе ненулевых κ^i не зависит от шага дискретизации пространства значений вектора z^x, результат можно распространить и на исходную задачу.

Утверждение. Решение задачи (3) включает (использует) не более m+1 различных векторов.

Таким образом, размерность пространства, по которому проводится поиск максимума, сокращается до $m\,(m+1)$. Применительно к задаче (2) это составляет 6, и задача легко решается численно.

Результаты

Предложенный метод позволяет найти зависимость $S_Q\left(\widetilde{F}_0\right)$ при любых параметрах n и N.

Однако наиболее нагляден и удобен для сравнения с другими подходами асимптотический случай: $\frac{N}{n} = M = \text{const}, \quad N \to \infty, \quad n \to \infty$. Рассматриваемое приближение вполне приемлемо уже при $n = 10$, при этом имеем лишь один входной параметр M.

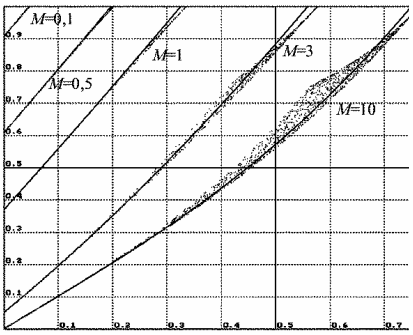
Для иллюстрации рассмотрим задачу классификации (k классов).

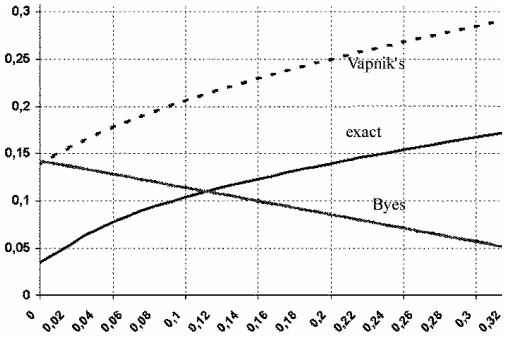
Функцией потерь будет: $L(y, y') = \begin{cases} 0, & y = y' \\ 1, & y \neq y' \end{cases}$.

Оценим $\hat{F}_x\left(p_x, \widetilde{F}_x^0\right)$ методом машинного моделирования. Для этого будем случайным образом генерировать ξ_x, то есть условное распределение $P_c[Y/x]$.

Условное распределение ξ_x задается k параметрами (вероятности для каждого образа). В соответствии с равномерным распределением на множестве параметров генерируем ξ_x, для которого вычисляем значения $F_x(p_x, \xi_x)$ и $\widetilde{F}_x(p_x, \xi_x)$ и полученную пару изображаем графически в виде точки.

В рассматриваемом асимптотическом приближении два входных параметра: N и p_x можно заменить одним: $\alpha = Np_x$.





Результаты моделирования при различных значениях параметра α приведены на левом рисунке (величины нормированы делением на $\frac{k-1}{k}$ – максимальную вероятность ошибки).

При $k = 2$ точки составляют сплошную линию. При $k > 2$ точки заполняют некоторую область, по которой также легко можно с достаточной точностью аппроксимировать $\hat{F}_x\left(p_x, \widetilde{F}_x^0\right)$.

Аналогичным образом, моделирование для оценки $\hat{F}_x\left(p_x, \widetilde{F}_x^0\right)$ можно провести и в случае прогнозирования вещественной переменной [4].

Теперь мы можем определить точность оценок Вапника–Червоненкиса для рассмотренного случая дискретного X, поскольку мы нашли точную зависимость средних риска от эмпирического риска для "наихудшей" стратегии природы.

Для $S_Q\left(\widetilde{F}_0\right)$ в [1] приводится оценка $\hat{S}'_V\left(\widetilde{F}_0\right) = \tau$, а также улучшенная оценка: $\hat{S}_V\left(\widetilde{F}_0\right) = \tau^2\left(1 + \sqrt{1 + \frac{2\widetilde{R}_0}{\tau^2}}\right)$, где τ асимптотически стремится к $\sqrt{\frac{\ln 2}{2M}}$.

На правом рисунке для $M = 5$ приведены зависимость $S_Q\left(\widetilde{F}_0\right)$ и ее оценка $\hat{S}_V\left(\widetilde{F}_0\right)$. График демонстрирует значительную завышенность последней.

Третья линия на графике (прямая) представляет вариант [5] искомой зависимости $\hat{S}_B\left(\widetilde{F}_0\right)=\frac{1-2\widetilde{F}_0}{M+2}$, получаемый Байесовским подходом, основанным на введении равномерного распределения на стратегиях природы [2,3,5].

Несколько неожиданным представляется тот факт, что $S_Q\left(\widetilde{F}_0\right)$ на значительном интервале оказывается существенно меньше $\hat{S}_B\left(\widetilde{F}_0\right)$, хотя Байесовский подход подразумевает усреднение по всем стратегиям природы, а изложенный метод — выбор наихудшей *c*.

Практическое использование результатов

Полученная зависимость может быть практически использована для оценки ожидаемого риска. Для этого достаточно к величине полученного эмпирического риска прибавить поправку $S_Q\left(\widetilde{F}_0\right)$.

Однако с прикладной точки зрения случай дискретного *X* и независимого принятия решения в каждой его точки имеет малый интерес.

Для использования полученного результата в реальных прикладных задачах предлагается следующий эвристический подход, основыванный на гипотезе, что зависимость $S_Q\left(\widetilde{F}_0\right)$ в общем случае будет близкой к найденной оценке, однозначно определяемой параметром *M*, которой можно оценить как $\frac{N}{\log_2 \mathbf{C}}$, где **C** – емкость класса решающих правил.

Однако данная оценка будет в точности адекватна только для алгоритма выбора решающей функции, осуществляющего полный перебор. Для алгоритмов направленного поиска фактическая емкость может оказаться меньше.

Оценить фактическую емкость используемого алгоритма построения решающей функции можно путем моделирования построения решающих функций при равномерном распределении в *D*.

Поскольку вероятность ошибки для любого правила при равномерном распределении известна (например, при распознавании двух образов она равна 0,5), то, оценив путем моделирования средний эмпирический риск, находим одну из точек (крайнюю правую) кривой $S_Q\left(\widetilde{F}_0\right)$, а значит, можем определить *M* и восстановить всю зависимость.

Работа выполнена при поддержке РФФИ, проект № 01-01-00839

СПИСОК ЛИТЕРАТУРЫ

1. В. Н. Вапник, А. Я. Червоненкис. Теория распознавания образов. М.: Наука, 1974. 415 с.
2. G. F. Hyghes. On the mean accuracy of statistical pattern recognizers // IEEE Trans. Inform. Theory. 1968. V. IT-14, N 1. P. 55–63.
3. Г. С. Лбов, Н. Г. Старцева. Логические решающие функции и вопросы статистической устойчивости решений. Новосибирск: Институт математики, 1999. 211 с.
4. V. M. Nedelko. An Asymptotic Estimate of the Quality of a Decision Function Based on Empirical Risk for the Case of a Discrete Variable. // Pattern Recognition and Image Analysis, Vol. 11, No. 1, 2001, pp. 69-72.
5. В. Б. Бериков. Об устойчивости алгоритмов распознавания в дискретной постановке. // Искусственный интеллект. Изд-во НАН Украины 2000, № 2. С. 5-8.