ОБ ЭФФЕКТИВНОСТИ ФУНКЦИОНАЛОВ ЭМПИРИЧЕСКОГО РИСКА И СКОЛЬЗЯЩЕГО ЭКЗАМЕНА КАК ОЦЕНОК ВЕРОЯТНОСТИ ОШИБОЧНОЙ КЛАССИФИКАЦИИ[1]

Виктор Неделько

Abstract: The goal of the paper is to estimate misclassification probability for decision function by training sample. The simple estimation presented shows how far Vapnik–Chervonenkis risk estimations are off in some special case. The existence of situations when an empirical risk appears to be better risk estimate in comparison with cross validation is investigated.

Keywords: pattern recognition, classification, statistical robustness, deciding functions, complexity, capacity, overtraining problem.

ACM Classification Keywords: G.3 Probability and statistics, G.1.6. Numerical analysis: Optimization; G.2.m. Discrete mathematics: miscellaneous.

Введение

Одним из наиболее часто используемых методов оценивания качества решающей функции является скользящий экзамен (cross-validation). Функционал скользящего экзамена, в отличие от эмпирического риска (доля неправильно классифицированных объектов обучающей выборки), является, как известно, несмещенной оценкой риска (вероятности ошибочной классификации).

Вместе с тем, оценка скользящего экзамена имеет относительно большую дисперсию, при этом достаточно точных универсальных оценок этой дисперсии до настоящего времени не найдено. Известные оценки имеют малую точность, и построенные на их основе интервальные оценки для риска оказываются, вообще говоря, не лучше, чем известные оценки Вапника-Червоненкиса (В-Ч) [Вапник, Червоненкис, 1974], основанные на эмпирическом риске.

В настоящей работе будет приведена простая оценка точности оценок В-Ч в некотором частном случае, а также будут исследованы ситуации, когда эмпирический риск является более эффективным функционалом, по сравнению со скользящим экзаменом. Под эффективностью здесь понимается точность оценивания риска.

Постановка задачи

Пусть X – пространство значений переменных, используемых для прогноза, а Y – пространство значений прогнозируемых переменных, и пусть C – множество всех вероятностных мер на $D = X \times Y$. Тогда элементом $c \in C$ будет $P_c[D]$. Здесь и далее квадратные скобки используются для указания множества, на σ-алгебре подмножеств которого задана мера.

[1] Работа выполнена при поддержке РФФИ, а также Лаврентьевского гранта СО РАН.

Решающей функцией назовем соответствие $f : X \to Y$ и введем для нее функцию потерь: $L : Y^2 \to [0, \infty)$.

Под риском будем понимать средние потери:

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

Пусть $v = \left\{ \left(x^i, y^i \right) \in D \;\middle|\; i = \overline{1, N} \right\}$ – случайная независимая выборка из распределения $P_c[D]$.

Эмпирический риск определим как средние потери на выборке:

$$\widetilde{R}(v, f) = \frac{1}{N} \sum_{i=1}^{N} L\left(y^i, f\left(x^i \right) \right).$$

Пусть $Q : \{v\} \to \Phi$ – алгоритм построения решающих функций, а $f_{Q,v} \in \Phi$ – функция, построенная по выборке v алгоритмом Q.

Оценкой скользящего экзамена называется величина

$$\breve{R}(v, Q) = \frac{1}{N} \sum_{i=1}^{N} L\left(y^i, f_{Q, v_i'}\left(x^i \right) \right),$$

где $v_i' = v \setminus \left\{ \left(x^i, y^i \right) \right\}$ – выборка, получаемая из v удалением i-го наблюдения.

Чтобы по полученному значению эмпирического риска можно было оценивать риск, необходимо так или иначе оценить вероятность уклонения данных величин:

$$\eta(c, \varepsilon) = \mathrm{P}\left(\left| \widetilde{R}\left(v, f_{Q,v} \right) - R\left(c, f_{Q,v} \right) \right| > \varepsilon \right).$$

Поскольку риск в среднем значительно больше эмпирического риска, возможно оценивать величину:

$$\eta'(c, \varepsilon) = \mathrm{P}\left(R\left(c, f_{Q,v} \right) > \widetilde{R}\left(v, f_{Q,v} \right) + \varepsilon \right).$$

Существенная проблема заключается в том, что выражения зависят от c – распределения, которое неизвестно. Решением может быть взятие супремума по всем распределениям и ориентирование таким образом на «наихудшее» распределение.

Кроме самих величин рисков нас будут интересовать их средние.

Обозначим: $F(c, Q) = \mathrm{E}R\left(c, f_{Q,v} \right)$, $\widetilde{F}(c, Q) = \mathrm{E}\widetilde{R}\left(c, f_{Q,v} \right)$, где математическое ожидание берется по всем выборкам объема N.

Все практически используемые алгоритмы построения решающих функций так или иначе минимизируют эмпирический риск, поэтому последний оказывается смещенной оценкой риска.

Введем функцию максимального смещения:

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

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

Классификация в дискретном пространстве

Будем рассматривать задачу классификации двух образов.

Пусть X дискретно, то есть $X = \{1,...,n\}$, и решающая функция минимизирует эмпирический риск независимо в каждой точке x.

При этом $Y = \{0, 1\}$, функцией потерь будет: $L(y, y') = \begin{cases} 0, & y = y' \\ 1, & y \neq y' \end{cases}$, а риском – вероятность ошибочной классификации.

Тогда вероятностная мера $c \in C$ задается набором вероятностей $c = \left\{ p_j^{\omega} = \mathrm{P}(x = j, y = \omega) \middle| \; j = \overline{1,n},\ \omega = \overline{0,1} \right\}$.

Чтобы проиллюстрировать степень точности оценок Вапника-Червоненкиса, рассмотрим так называемый «детерминистский» случай, когда в классе решающих функций всегда находится решение, которое классифицирует выборку точно, т. е. $\widetilde{R} = 0$.

При этом будем рассматривать асимптотический случай: $\frac{N}{n} = M = \text{const}$, $N \to \infty$, $n \to \infty$. Следует подчеркнуть, что мы рассматриваем «асимптотику малых выборок», то есть условия, близкие к условиям подавляющего большинства реальных задач.

Для начала приведем вывод асимптотических выражений для оценок В-Ч, когда эмпирический риск равен нулю.

Вероятность уклонения для риска находится как:

$$P\left(R - \widetilde{R} \geq \varepsilon \right) = P\left(\widetilde{R} = 0 / R = \varepsilon \right) = (1 - \varepsilon)^N .$$

Для равномерного уклонения имеем оценку:

$$P\left(\sup_{f \in \Phi} \left| R - \widetilde{R} \right| \geq \varepsilon \right) < |\Phi| (1 - \varepsilon)^N .$$

Приравняем правую часть величине выбранного уровня значимости η:

$$\ln|\Phi| + N \ln(1 - \varepsilon) = \ln \eta .$$

Легко заметить, что в асимптотике больших N слагаемым $\ln \eta$ можно пренебречь.

Теперь, учитывая, что $|\Phi| = 2^{n\left(1 - e^{-M}\right)}$ и $\ln(1 - \varepsilon) \approx -\varepsilon$ получаем:

$$S'_V(0) = \varepsilon = \frac{\left(1 - e^{-M}\right) \ln 2}{M} .$$

Множитель $1 - e^{-M}$ есть вероятность ненулевого числа исходов при распределении Пуассона, или средняя доля значений переменной X, с ненулевым числом выборочных точек. Данная поправка отражает тот факт, что число решающих функций, различимых на выборке, определяется именно числом «непустых» значений.

Точное значение смещения риска при $\widetilde{F}_0 = 0$ составляет:

$$S_Q(0) = \begin{cases} \frac{1}{2}e^{-M}, & M \le 1 \\ \frac{1}{2Me}, & M \ge 1 \end{cases}.$$

Данный результат может быть получен следующим образом.

Во-первых, заметим, что ожидание эмпирического риска может быть равно нулю, только если распределения классов не пересекаются, т. е. $p_j^0 p_j^1 = 0,\ j = \overline{1,n}$.

Во-вторых, ошибочное решение возможно только для тех значений X, в которые не попало ни одной выборочной точки, причем вероятность такого события для рассматриваемого асимптотического приближения находится из распределения Пуассона, а вероятность ошибки при условии попадания в такое значение равна ½. Таким образом, вероятность ошибки всецело (считаем, что априорные вероятности классов равны) определяется априорным распределением в X.

Для нахождения распределения, при котором ожидаемая вероятность ошибки максимальна, теперь воспользуемся фактом, доказанным в [Неделько, 2003], а именно, тем, что максимум смещения риска достигается на кусочно-постоянном распределении, с двумя (для рассматриваемого случая) областями постоянства. Более того, исследование показывает, что одна из областей имеет минимально возможный размер.

Таким образом, для искомого распределения в одной из точек X сосредоточена вероятность $1 - \alpha$, а по остальным равномерно распределена вероятность $0 < \alpha \le 1$.

Тогда (после предельного перехода) ожидаемая вероятность ошибки будет $\frac{1}{2}\alpha e^{-\alpha M}$.

Максимум данного выражения достигается при $\alpha = \min\left(\frac{1}{M}, 1\right)$, откуда и получаем искомое смещение риска.

Отношение $\frac{S'_V(0)}{S_Q(0)} = \frac{2Me\left(1 - e^{-M}\right)\ln 2}{M} \xrightarrow[M\to\infty]{} 2e\ln 2 \approx 3{,}77$ показывает степень погрешности оценок В-Ч, которая, в основном, обусловлена приближением вероятности суммы суммой вероятностей, а также заменой энтропии на емкость.

Вклад последнего фактора можно оценить. Для этого заметим, что для найденного «худшего» распределения доля «непустых» значений переменной X при $M > 1$ на самом деле равна $1 - e^{-1}$, откуда получаем: $\frac{S'_V(0)}{S_Q(0)} = 2(e-1)\ln 2 \approx 2{,}38$.

Сравнение эффективности функционалов

Исчерпывающее сравнение эффективности функционалов в настоящее время не представляется возможным, поскольку на их основе пока не построены оптимальные оценки доверительного интервала для вероятности ошибки. Поэтому будем проводить сравнение на некоторых частных примерах. Нашей целью будет показать, что существуют такие ситуации, когда функционал эмпирического риска оказывается эффективнее. Для этого нам достаточно рассмотреть ряд весьма упрощенных модельных задач.

Пусть в дискретной задаче классификации, сформулированной в предыдущем разделе, распределение в

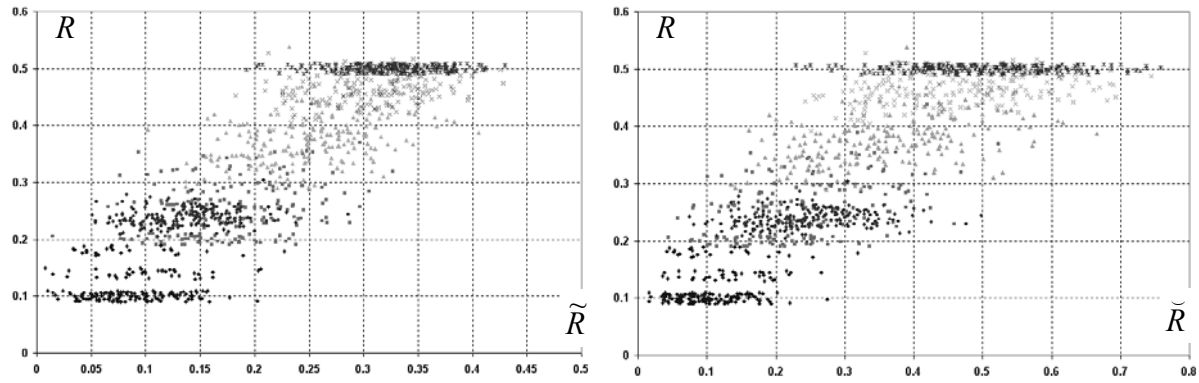

Рис. 1. Точки изображают полученные моделированием случайных выборок пары значений риска и эмпирического риска (левый график), а также риска и функционала скользящего экзамена (правый график). Разные типы маркеров соответствуют значениям параметра ξ: 0,1, 0,2, 0,3, 0,4, 0,5. Объем выборок N = 50, мощность пространства n = 10.

пространстве D задается одним параметром $0 \le \xi \le 0{,}5$, а именно:

$$P(x = j, y = 0) = \frac{\xi}{n},\ \ P(x = j, y = 1) = \frac{1-\xi}{n},\ \ j = \overline{1,n}.$$

Параметр ξ определяет так называемый байесовский уровень ошибки, т. е. вероятность ошибочной классификации при оптимальной решающей функции.

Решающая функция минимизирует эмпирический риск, т. е.

$$f(x) = \arg\max_{y \in \{0,1\}} \left| \left\{ \left(x^i, y^i\right) \in v \,\middle|\, x^i = x,\ y^i = y \right\} \right|.$$

В случае, если при заданном x число выборочных точек обоих классов одинаково, в качестве решения принимаются с равной вероятностью 0 и 1.

На рис. 1 приведены результаты моделирования случайных выборок. Видно, что дисперсия функционала скользящего экзамена больше дисперсии эмпирического риска, однако визуально сложно оценить, какой из функционалов позволяет более точно прогнозировать риск.

На практике распределение c неизвестно, поэтому для сравнения эффективности функционалов нужно на их основе построить наилучшие оценки доверительных интервалов. Для рассмотренной дискретной постановки эта задача, по-видимому, решаема, однако, достаточно нетривиальна (см. далее), и к настоящему времени решение получить не удалось.

Однако, поставленная нами цель позволяет сделать упрощение и предположить, что c определяется одним параметром ξ и, более того, само случайно, задав при этом некоторое распределение на множестве значений параметра ξ.

Тогда определено полное совместное распределение величин R, $\widetilde{R}$ и $\breve{R}$.

Критерием качества (эффективности) функционала $\widetilde{R}$ примем

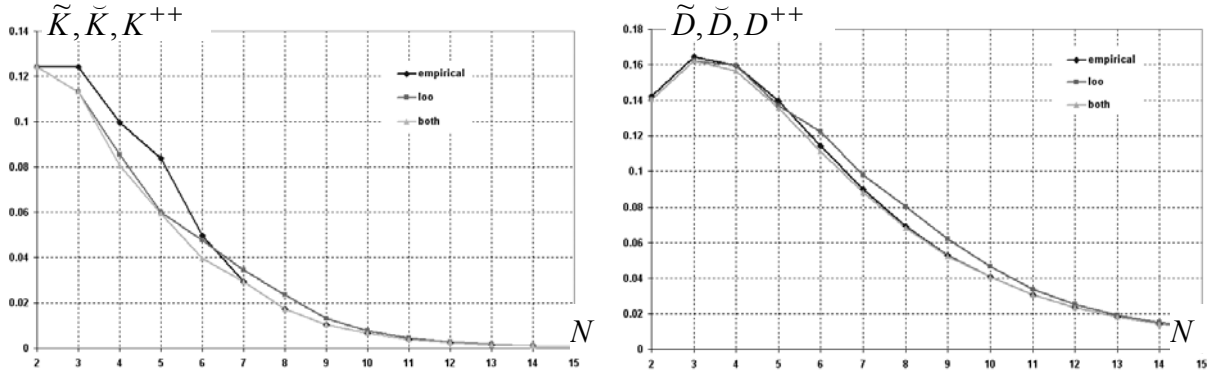

Рис. 2. Сравнение эффективности функционалов эмпирического риска и скользящего экзамена при различном объеме выборки, n = 3.

$$\widetilde{K} = \int \left(\min_{R^* \in [0,1]} \int \left| R - R^* \right| dP\left(R/\widetilde{R}\right) \right) dP\left(\widetilde{R}\right).$$

Аналогично

$$\breve{K} = \int \left(\min_{R^* \in [0,1]} \int \left| R - R^* \right| dP\left(R/\breve{R}\right) \right) dP\left(\breve{R}\right).$$

Содержательно данные критерии отражают среднюю точность (в смысле модуля отклонения) прогнозирования риска при использовании соответствующего выборочного функционала.

Можно также измерять погрешность через среднеквадратичное уклонение:

$$\widetilde{D} = \sqrt{\int \left(\int \left(R - E_{\widetilde{R}} R \right)^2 \, dP\left(R/\widetilde{R}\right) \right) dP\left(\widetilde{R}\right)},$$

где $E_{\widetilde{R}} R$ – условное математическое ожидание риска при заданной величине эмпирического риска. Аналогично вводится $\breve{D}$ – среднеквадратичный критерий качества для скользящего экзамена.

На рис. 2 изображены зависимости качества функционалов эмпирического риска и скользящего экзамена от объема выборки при n = 3. При этом параметр ξ равновероятно принимает значения 0 и 0,5. Видно, что различие эффективности функционалов в рассмотренном примере невелико, однако, существуют значения параметров, при которых функционал эмпирического риска достоверно эффективнее функционала скользящего экзамена.

Совместное использование функционалов

Оценивать риск можно и при одновременном использовании обоих функционалов.

$$K^{++} = \int \left(\min_{R^* \in [0,1]} \int \left| R - R^* \right| dP\left(R/\widetilde{R}, \breve{R}\right) \right) dP\left(\widetilde{R}, \breve{R}\right).$$

В этом случае эффективность оценивания риска может быть несколько выше, чем при использовании любого из функционалов в отдельности. Это демонстрируется третьей (наиболее светлой) кривой на рис. 2.

Более общим случаем метода скользящего экзамена является отделение для контроля не по одному объекту выборки, а выбор всех сочетаний из k объектов.

Естественно ожидать, что более точная оценка риска может быть получена при использовании полного набора функционалов скользящего экзамена, включая все допустимые значения параметра k.

Построение доверительного интервала

Чтобы построить доверительный интервал для риска на основе некоторого функционала, например, эмпирического риска, нужно выбрать критическое множество из множества пар $\left(\widetilde{R}, R\right)$.

Зададим критическое множество в виде

$$\left\{\left(\widetilde{R}, R\right) \middle|\ R > \hat{R}_\eta\left(\widetilde{R}\right)\right\},$$

тогда доверительным интервалом будет интервал $[\,0, \hat{R}_\eta\left(\widetilde{R}\right)]$.

Функция $\hat{R}_\eta\left(\widetilde{R}\right)$ определяется из условий:

$$\forall c,\ \ P\left(R > \hat{R}_\eta\left(\widetilde{R}\right)\right) \le \eta\,,$$

$$\int_0^{\widetilde{R}_{\max}} \hat{R}_\eta\left(\widetilde{R}\right) d\widetilde{R} \to \min\,.$$

Первое из условий отражает требование, что вероятность попадания в критическую область не должна превышать заданного уровня значимости. Второе условие отражает предпочтительность по возможности меньших величин прогнозируемого риска. Точное выражение данного условия может корректироваться в зависимости от практических требований к прогнозу.

Точное решение возникающей оптимизационной задачи до сих пор не найдено, но известно приближенное решение, предложенное Вапником и Червоненкисом, которые, однако, не формулировали оптимизационную задачу в представленном виде, а подбирали $\hat{R}_\eta\left(\widetilde{R}\right)$ эмпирически.

Заключение

В работе рассмотрен ряд случаев, когда функционал эмпирического риска эффективнее скользящего экзамена. Само по себе существование таких случаев представляется интересным и говорит о том, что при условии точной оценки смещения эмпирический риск может конкурировать с несмещенной оценкой скользящего экзамена.

Обнаруженное преимущество эмпирического риска наблюдалось лишь в некоторых случаях и было несущественным, поэтому на практике все же более предпочтительным остается использование (при наличии такой возможности) функционала скользящего экзамена.

Задача построения оптимального функционала оценки качества в настоящее время остается открытой.

Литература

[Вапник, Червоненкис, 1974] Вапник В. Н., Червоненкис А. Я. Теория распознавания образов. М.: Наука, 1974. 415 с.

[Лбов, Старцева, 1999] Г.С. Лбов, Н.Г. Старцева. Логические решающие функции и вопросы статистической устойчивости решений. Институт математики СО РАН, Новосибирск, 1999, 211 с.

[Неделько, 2003] V. M. Nedelko. Estimating a Quality of Decision Function by Empirical Risk // LNAI 2734. Machine Learning and Data Mining in Pattern Recognition. Third International Conference, MLDM 2003, Leipzig. Proceedings. Springer-Verlag. pp. 182–187.

[Неделько, 2004] Неделько В. М. Оценивание смещения эмпирического риска для линейных классификаторов. // Таврический вестник информатики и математики. Изд-во НАН Украины 2004, № 1. С. 47–53.

Информация об авторе

Виктор Михайлович Неделько – с.н.с. лаборатории Анализа данных Института математики СО РАН, 660090, пр-т Коптюга, 4, Новосибирск, Новосибирский госуд. Университет, Россия, e-mail: nedelko@math.nsc.ru