УДК 519.9

Об интервальном оценивании риска для решающей функции

© Неделько В.М.

Институт математики СО РАН
Лаб. Анализа данных
пр-т Коптюга, 4, г. Новосибирск, 630090, Россия

e-mail: *nedelko@math.nsc.ru*

Abstract. The problem of statistical decisions risk estimates construction by the absence of any information on the probabilistic distribution is considered. A method of empirical confidence interval construction via selection of a finite set of distributions is proposed. The method was probed on the task of classification by the nearest neighbour[3].

Введение

Рассмотрим следующую достаточно общую постановку задачи построения решающей функции. Пусть Γ – некоторая (генеральная) совокупность объектов, и $\nu \subseteq \Gamma$ – выборка из нее. Требуется сопоставить выборке некоторую $f \in \Phi$ – решающую функцию так, чтобы она в определенном смысле характеризовала генеральную совокупность. Последнее означает, что задана так называемая функция потерь $L : \Gamma \times \Phi \to \Re$, и необходимо выбрать отображение $Q : V \to \Phi$, которое бы минимизировало потери. Отображение Q будем называть алгоритмом построения решающей функции. Здесь V – множество всех выборок заданного объема.

Первая проблема состоит в том, что на основе заданной функции потерь нужно сформировать некоторый функционал качества, характеризующий алгоритм на всей генеральной совокупности, что требует наличия оснований для переноса свойств ν на Γ. Одно из решений заключается в предположении существования на Γ некоторой вероятностной меры $c \in C$, в соответствии с которой выборка ν генерируется случайным образом. Говоря о вероятностной мере, мы подразумеваем также задание подходящего вероятностного пространства.

Теперь, чтобы охарактеризовать качество решения на всей Γ, можно на основе функции потерь определить функционал риска $R(c, f)$, например, как средние (ожидаемые) потери.

Заметим, что это не единственный подход к проблеме экстраполяции свойств ν на Γ, одна из альтернатив предложена в [2].

Для оценивания качества решения, то есть функционала риска, рассмотрим возможность построения доверительных интервалов.

Определение 1. Функция $K : V \to I \subset 2^{\Re}$ называется *доверительным интервалом*, если

$$\forall c \in C,\ P\left(R\left(c, g(\nu)\right) \in K(\nu)\right) \geqslant \eta \in (0,\, 1).$$

[3]Работа выполнена при поддержке РФФИ, проекты № 07-01-00331-а и № 08-01-00944-а.

Впервые нетривиальные доверительные интервалы для риска в задаче классификации построили Вапник и Червоненкис [1]. При этом важен факт получения оценок без каких-либо предположений о виде распределений.

Если рассматривать статистические задачи вообще, то первым подобным результатом, видимо, является теорема Гливенко. Заметим, что в формулировках этой теоремы не отражен факт равномерной сходимости по c, хотя из доказательства он следует.

Оценки Вапника-Червоненкиса используют характеристики класса Ф, а именно емкостную меру его сложности. Альтернативой емкостным характеристикам может служить колмогоровская сложность [3].

Известно, что оценки Вапника-Червоненкиса являются сильно завышенными, однако в некоторых случаях можно получить более приемлемую точность [4]. Очевидно, что лучшие оценки можно получить [5], рассматривая свойства самого алгоритма Q, а не класса Ф. Такие оценки могут быть нетривиальными, даже если класс Ф имеет бесконечную емкость.

Построение доверительного интервала для риска требует оценивания вероятности по всем возможным распределениям. Поскольку известные аналитические оценки такого рода имеют большую погрешность, оправдано построение эмпирических оценок путем статистического моделирования на широком классе распределений. В работе представлены результаты применения данного подхода к оцениванию риска для метода классификации по ближайшему прецеденту.

1. Постановка задачи построения решающей функции

Пусть X – пространство значений переменных, используемых для прогноза, а Y – пространство значений прогнозируемых переменных, и пусть C – множество всех вероятностных мер на заданной σ-алгебре подмножеств множества $D = X \times Y$. При каждом $c \in C$ имеем вероятностное пространство: $\langle D, B, P_c \rangle$, где B – σ-алгебра, $P_c[D]$ – вероятностная мера. Параметр с будем называть *стратегией природы*.

Решающей функцией называется соответствие $f : X \to Y$.

Качество принятого решения оценивается заданной функцией потерь $L : Y^2 \to [0, \infty)$. Под риском будем понимать средние потери:

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

Пусть $\nu = \{(x^i, y^i) \in D \,|\, i = \overline{1, N}\}$ – случайная независимая выборка из распределения $P_c[D]$. Эмпирический риск определим как средние потери на выборке:

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

Заметим, что значение риска зависит от стратегии природы c – распределения, которое неизвестно. Функционал скользящего экзамена определяется как:

$$\breve{R}(\nu, Q) = \frac{1}{N}\sum_{i=1}^{N} L(y^i, f_{Q,\nu_i'}(x^i)),$$

где $\nu_i' = \nu\backslash\{(x^i, y^i)\}$ – выборка, получаемая из ν удалением i-го наблюдения, $Q : \{\nu\} \to \Phi$ – алгоритм построения решающих функций, $f_{Q,\nu}$ – функция, построенная по выборке ν алгоритмом Q, Φ – заданный класс решающих функций.

Задача построения решающей функции заключается в выборе подходящего алгоритма Q и в оценивании риска принятого решения.

2. Методы оценивания риска

Доверительный интервал для R будем задавать в виде $[0, \hat{R}(\nu)]$. Здесь мы ограничиваемся односторонними оценками, поскольку на практике для риска важны именно оценки сверху. Таким образом, в данном случае построение доверительного интервала эквивалентно выбору функции $\hat{R}(\nu)$, которую будем называть оценочной функцией или просто оценкой (риска).

При этом должно выполняться условие:

$$\forall c, P_c\big(R \leqslant \hat{R}(\nu)\big) \geqslant \eta,$$

где η – заданная доверительная вероятность.

При построении оценок риска первая проблема, которую нужно решить, это сравнение качества различных оценок.

Можно положить, что задан функционал качества $\Upsilon\big(F_{c,\hat{R}}(\cdot)\big)$, где $F_{c,\hat{R}}(\cdot)$ – функция распределения оценки $\hat{R}(\nu)$. Выбор данного функционала, так же как и выбор функции потерь, определяется практическими соображениями. Простейшим вариантом такого функционала является математическое ожидание.

При фиксированной стратегия природы c функционал Υ позволяет сравнивать качество оценок риска и находить оптимальную оценку.

Однако на практике распределение c неизвестно, а оценки, оптимальной при всех распределениях, может не существовать. В этом случае естественным является поиск множества Парето недоминируемых оценок.

Известные на данный момент оценки риска (напр. [1]) строятся не как функции непосредственно выборки, а через композицию $\hat{R}(\nu) = \hat{R}_e\big(\dot{R}(\nu)\big)$, то есть как функции значений некоторого эмпирического функционала $\dot{R}$, в качестве которого обычно выступает эмпирический риск или скользящий экзамен.

Эмпирический функционал здесь выступает в роли точечной оценки риска, на основе которой строится интервальная оценка.

Актуальной является задача исследования эффективности функционалов эмпирического риска и в различных прикладных задачах. При этом под эффективностью понимается, насколько хорошая интервальная оценка риска может быть построена на основе данного функционала.

Существует очевидный класс задач, в которых скользящий экзамен эффективнее — это случаи использования алгоритмов бесконечной емкости, для которых эмпирический риск всегда равен нулю. При этом для многих алгоритмов бесконечной емкости возможно построение оценки риска на основе функционала скользящего экзамена, что будет проиллюстрировано ниже. Однако для других алгоритмов предпочтительность скользящего экзамена уже не очевидна. Один из доводов в пользу эмпирического риска состоит в том, что он характеризует именно построенную решающую функцию f, в то время как скользящий экзамен характеризует алгоритм Q в целом. Привлекательной выглядит идея одновременного использования обоих функционалов.

Определение 2. Оценочную функцию $\hat{R}(\nu)$ назовем *согласованной* с эмпирическим функционалом $\dot{R}$, если для выборок ν_1 и ν_2 одинакового объема

$$\dot{R}(\nu_1) > \dot{R}(\nu_1) \Rightarrow \hat{R}(\nu_1) \geq \hat{R}(\nu_2).$$

Достаточно естественным представляется ограничиться рассмотрением только таких оценочных функций, которые согласованы с функционалами эмпирического риска и скользящего экзамена. Это означает, что оценка вероятности ошибки не должна убывать при увеличении значения эмпирического функционала.

Данное условие позволяет резко сузить пространство поиска при нахождении Парето-оптимальных оценочных функций.

3. Построение эмпирической оценки риска

Недоминируемость оценочной функции является безусловно желательным, но трудно проверяемым свойством. Более того, на практике оказывается проблематичным даже оценивание доверительной вероятности для заданной оценочной функции, поскольку это подразумевает взятие супремума по всем распределениям. Класс распределений иногда может быть ограничен некоторым параметрическим семейством, однако при отсутствии априорной информации единственным ограничением становится измеримость функции потерь, что на практике обычно означает допустимость любых вероятностных мер на σ-алгебре борелевских множеств пространства переменных.

Задача построения точных аналитических оценок доверительной вероятности в настоящее время не решена, поэтому на практике оправданным является построение эмпирических оценок. Под эмпирической оценкой понимается величина, полученная оцениванием максимальной доверительной вероятности по некоторому эвристически выбранному множеству распределений. Если это множество выбрано достаточно «богатым», то естественно ожидать, что полученная оценка будет близка к истинной. При этом не предполагается оценивание точности данного эмпирического приближения, однако, возможность доверия такому подходу может быть аргументирована следующим соображением. Если целенаправленным эвристическим поиском не удалось построить распределения, при котором доверительная вероятность была бы меньше заданной величины, то можно ожидать, что и в реальной задаче распределение окажется таким, что оценка останется справедливой.

4. Иллюстративный пример

Приведем пример использования данного метода для построения доверительного интервала для риска в задаче классификации методом ближайшего соседа.

Пусть $X_1, ..., X_n$ – непрерывные переменные, а $Y \in \{1, 2\}$ – номер класса.

Зададим семейство распределений, которые задаются плотностью вероятности $\rho(x, y) = \varphi(x) \cdot P\left({}^{y}/_{x}\right)$, где $\varphi(x)$ – равномерное распределение в гиперкубе $[0, 1]^n$, а $P\left({}^{y}/_{x}\right) = \begin{cases} g(x), & y = 1 \\ 1 - g(x), & y = 2 \end{cases}$ – функция условной вероятности. Функция $g(x) = P\left({}^{y=1}/_{x}\right)$ задается как $g(x) = \begin{cases} \lambda, & \psi(x) = 1 \\ 1 - \lambda, & \psi(x) = -1 \end{cases}$, $\psi(x) = \prod\limits_{j=1}^{n} \Psi(kx_j)$, $x = (x_1, ..., x_n)$. Функция $\Psi(\cdot)$ принимает значение 1, если целая часть аргумента является четным числом, и значение -1 в противном случае.

В приведенной модели гиперкуб $[0, 1]^n$ разбит на k^n областей, которым в «шахматном» порядке приписаны классы. Параметр λ соответствует байсовскому риску (вероятности ошибочной классификации для наилучшего правила).

Оценочную функцию будем строить как функцию от числа ошибок скользящего экзамена.

Результаты моделирования при $N = 50$ приведены на рис. 1. Точки (маркеры) на графике отображают пары значений $(\breve{R}, R)$ для некоторой случайной выборки. Разные виды маркеров соответствуют различным распределениям, параметры которых отражены в легенде, где $a = 100\lambda$, $n = 1$. Помимо отраженных на графике для построения эмпирического доверительного интервала были использованы еще 20 распределений с различиными параметрами.

Построены две оценочных функции: градиентная и линейная (серая и черная кривые на графике). Доверительная вероятность $\eta = 0,9$.

Градиентный алгоритм работает итеративно. В качестве начального приближения оценочной кривой берется $\hat{R}(\nu) \equiv 0$. Далее на каждом шаге находится распределение, при котором вероятность выхода за оценочную кривую максимальна. Затем функция $\hat{R}(\nu)$ увеличивается на фиксированную малую величину в точке, где это изменение максимизирует доверительную вероятность для этого распределения. Итерации прекращаются, когда текущая доверительная вероятность станет не меньше порогового значения η.

Полученная градиентным методом оценка оказалась достаточно близка к линейной, что говорит о разумности построения оценки в классе линейных функций. Чтобы обеспечить единственность решения, потребуем, чтобы оценка максимизировала площадь над ее графиком. Оценочную прямую удобно задавать двумя точками: $(0,\ R_{min})$ и $(\breve{R}_{max},\ 0,5)$. В рассмотренном примере получены значения: $R_{min} = 0,035$, $\breve{R}_{max} = 0,365$. Первый параметр есть минимум оценки риска, который достигается при нулевом числе ошибок на скользящем экзамене. Второй параметр есть значение доли ошибок на скользящем экзамене, начиная с которого оценка риска равна $0,5$.

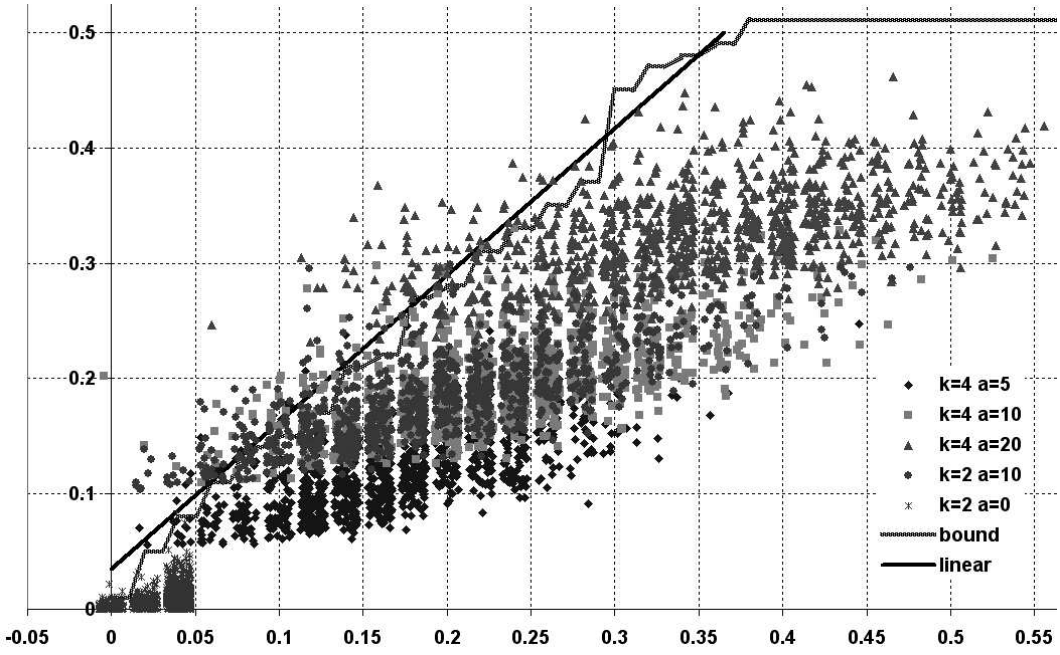

Рис. 1. Зависимость риска от ошибки скользящего контроля.

5. Общий вид доверительного предиката

Понятие доверительного интервала можно обобщить введением понятия доверительного предиката.

Определение 3. Предикат $T(c, \nu)$ называется *доверительным предикатом*, если

$$\forall c \in C,\ P\left(T(c,\nu)\right) \geqslant \eta \in (0,\,1).$$

В случае построения доверительных интервалов $T(c, \nu) = (R\left(c, g(\nu)\right) \in K(\nu))$. В общем случае предикату соответствует некоторое подмножество в $C \times V$.

Доверительный предикат можно использовать для проверки статистических гипотез. Для этого введем функцию $\Theta(\nu) = \{c \mid T(c, \nu)\}$. Данная функция каждой выборке сопоставляет множество согласующихся с ней гипотез о распределении. Приведенная схема используется классическими методами статистики для оценивания параметров распределений в случае, когда C представляет собой параметрическое семейство. Однако нет принципиальных препятствий для использования этого подхода и в непараметрическом случае, где он может являться альтернативой критериям согласия. Примером может служить известная теорема Гливенко, если ее вместе с сопутствующими результатами переформулировать подходящим образом. А именно,

$$\forall c \in C,\ P\left(\sup_x |\widetilde{F}(x) - F(x)| < \varepsilon(\eta))\right) \geqslant \eta,$$

где $F(x)$ – функция распределения случайной величины X, а $\widetilde{F}(x)$ – эмпирическая функция распределения.

При построении доверительного предиката представляется перспективным использование эмпирических методов оценивания доверительной вероятности.

Заключение

В работе предложен метод оценивания риска для решающей функции посредством построения эмпирического доверительного интервала. Эффективность метода проиллюстрирована на примере задачи классификации по ближайшему прецеденту. В отличие от сложностных оценок Вапника–Червоненкиса метод построения эмпирических доверительных интервалов применим также для алгоритмов, использующих классы решающих функций бесконечной емкости.

Список литературы

1. *Вапник В.Н., Червоненкис А.Я.* Теория распознавания образов. М.: Наука, 1974. 415 с.
2. *Воронцов К.В.* Слабая вероятностная аксиоматика и надежность эмпирических предсказаний. // Доклады 13-й всероссийской конференции «Математические методы распознавания образов». М. 2007. С. 21–24.
3. *Донской В.И.* Колмогоровская сложность классов общерекурсивных функций с ограниченной емкостью. // Таврический вестник информатики и математики. НАН Украины. 2005, № 1. С. 25–34.
4. *Неделько В.М.* Об оценивании вероятности ошибочной классификации. // Искусственный интеллект. Изд-во НАН Украины, 2006, № 2. С. 197–200.
5. *Неделько В.М.* Об эффективности эмпирических функционалов качества решающей функции. // Доклады 13-й всероссийской конференции «Математические методы распознавания образов». М. 2007. С. 47–49.

Статья поступила в редакцию 01.05.2008