МОДЕЛИРОВАНИЕ И ОПТИМИЗАЦИЯ РОБАСТНЫХ ОЦЕНОК ФУНКЦИЙ ПО НАБЛЮДЕНИЯМ

Е. С. Кирик
Институт вычислительного моделирования СО РАН, Красноярск, Россия
e-mail: kirik@icm.krasn.ru

В работе рассматривается цензорный подход к построению и оптимизации робастных оценок функций по наблюдениям и "ремонту" данных. В качестве модели восстанавливаемой неизвестной зависимости принимается непараметрическая оценка регрессии. С целью оптимизации полученной оценки предлагается критерий качества "очистки" выборки.

Введение

Как правило, в обрабатываемых экспериментальных данных в силу различных причин на фоне "типичных", "средних" наблюдений, представляющих собой выборку из некоторой единой генеральной совокупности, содержатся значительные отклонения — "выбросы", "промахи". Среднее количество грубых измерений в данных колеблется около 10–15 % [13]. Наличие последних приводит к нарушению условий оптимальности для классических процедур оценивания неизвестных зависимостей по наблюдениям. Задачей теории робастного оценивания является разработка таких процедур, которые позволяют получать высокое качество оценок в этих условиях, но уступают в качестве классическим процедурам при выполнении условий их оптимальности [1]. Термин "робастный" в указанном смысле впервые был введен Боксом в 1953 г. [7].

Чаще всего априорная информация о вероятностных характеристиках "загрязнений" экспериментальных данных и о возможном виде восстанавливаемой зависимости отсутствует. Так же априори неизвестен ни сам факт наличия выбросов, ни наблюдения их содержащие. Поэтому моделировать оценку, устойчивую к выбросам, предпочтительнее в предположении довольно слабых ограничений на саму неизвестную функцию, а так же на данные ее представляющие. Для этих целей наиболее подходит непараметрический подход к анализу данных. С одной стороны здесь в качестве модели неизвестной зависимости рассматривается регрессия, в которой вероятностные характеристики случайных величин заменены их ядерными оценками [10], и следовательно единственным ограничением на функцию является требование ее однозначности. С другой, свойства самих непараметрических оценок и инструмент непараметрической статистики позволяют моделировать робастные оценки цензорного типа и строить критерии оптимальности для этих оценок. Преимущество оценок цензорного типа перед оценками обладающими свойством сглаживания (в первую очередь это непараметрические оценки основанные на минимаксном подходе Хьюбера [9]: [3, 6] и др.) состоит в исключении, точек представляющих выбросы из рассмотрения и, соответственно, исключении влияния последних на оценку. Однако известные цензорные подходы ([8, 5, 6, 12, 13] и др.) к анализу данных не предполагают этапа проверки качества "очистки" выборки ввиду неформализованности критерия оптимальности, кроме того наличие выбросов в данных является обязательным требованием. Для предлагаемого непараметрического подхода последнее условие не является обязательным, что позволяет получить качество робастной оценки в отсутствии выбросов не уступающее качеству классической непараметрической процедуры оценивания неизвестных функций по наблюдениям.

1. Постановка задачи

Пусть дана обучающая выборка $V = \{x_i, y_i\}, i = \overline{1,n}$ — статистическая выборка независимых наблюдений (x, y) с помехами случайной величины, распределенной с неизвестной плотностью вероятности $p(x, y)$ и $p(x) > 0 \quad \forall x \in x$. Считаем, что $x = (x^1, ..., X^l)$, помехи имеют нулевое математическое ожидание, вид нелинейной стохастической зависимости $y = f(x)$ однозначный. Предполагаем, что выбросы имеют симметричное распределение и могут составлять до 15 % от объема обучающей выборки. Для удобства данные считаются нормированными и центрированными.

*Работа поддержана Красноярским краевым фондом науки, № 10F123N и № 12G188.

Требуется построить непараметрическую робастную оценку неизвестной зависимости $y = f(x)$ по ее наблюдениям $V = \{x_i, y_i\}, i = \overline{1,n}$.

2. Робастная оценка регрессии

В общем случае $(x_i = (x_i^1, ..., x_i^l))$ сходящаяся [4] непараметрическая оценка условного математического ожидания $\hat{y} = \hat{f}(x) = M(y/x) = \int yp(y/x)\,dy$ (или регрессии) имеет вид

$$y_n(x) = \frac{\sum\limits_{i=1}^{n} y_i \prod\limits_{j=1}^{l} \Phi(\frac{x^j - x_i^j}{C_n})}{\sum\limits_{i=1}^{n} \prod\limits_{j=1}^{l} \Phi(\frac{x^j - x_i^j}{C_n})}. \quad (1)$$

Оценка (1) получается из (??) подстановкой в нее оценок плотностей типа Розенблата-Парзена ([11], [10]) с учетом условия самовоспроизводимости $\Phi(\cdot)$

$$\frac{1}{C_n} \int\limits_{\Omega(y)} y\Phi(\frac{y^j - y_i^j}{C_n})\,dy = y_i, i = \overline{1,n}, j = \overline{1,l}. \quad (2)$$

$\Phi(\cdot)$ — финитная колоколообразная интегрируемая с квадратом функция, удовлетворяющая условиям

$$0 < \Phi(z) < \infty, \forall z \in \Omega(z);\ \frac{1}{C_n} \int\limits_{\Omega(x)} \Phi(\frac{x - x_i}{C_n})\,dx = 1;$$

$$\lim_{n\to\infty} \frac{1}{C_n}\Phi(\frac{x - x_i}{C_n}) = \delta(x - x_i); \quad (3)$$

C_n — параметр размытости такой, что

$$C_n > 0; \quad \lim_{n\to\infty} C_n = 0; \quad \lim_{n\to\infty} nC_n^k = \infty. \quad (4)$$

Последний является неизвестным параметром в (1), подлежащим определению. Оптимальный параметр размытости C_n соответствует минимуму квадратичного критерия оптимальности

$$\omega^2(C_n) = \sum_{i=1}^{n} \big(y(x_i) - \tilde{y}_n(x_i, C_n)\big)^2 \to \min_{C_n} \quad (5)$$

и находится в ходе скользящего экзамена на обучающей выборке.

Оценка (1) будучи точечной взвешенной оценкой, очевидно, чувствительна к наличию выбросов в данных. Предварительный анализ, предполагающий исследование последних и исключение "подозрительных" элементов из рассмотрения позволяет перейти от непараметрической оценки регрессии (1) к ее робастному аналогу.

Информативным с точки зрения выделения "подозрительных" элементов выборки представляется исследование невязок $\epsilon_i = y_i - y_n(x_i), i = \overline{1,n}$ и функций от них ($y_n(x_i), i = \overline{1,n}$ — оценки (1) элементов выборки, вычисленные в режиме скользящего экзамена при оптимальном C_n в смысле критерия (5)). Упорядоченные по возрастанию величины $\epsilon_i, i = \overline{1,n}$ образуют вариационный ряд

$$\epsilon^1 \le \epsilon^2 \le ... \le \epsilon^n, \quad (6)$$

где $\epsilon^1 = \min \epsilon_i$, $\epsilon^n = \max \epsilon_i, i = \overline{1,n}$.

На рис. 1 видно, что большей частью они расположены компактно на числовой оси, и лишь некоторые лежат в отдалении от общей массы. Такое распределение невязок обусловлено свойством локальной аппроксимации оценки (1) и свидетельствует о наличии выбросов в выборке. Оценки последних есть взвешенное среднее "хороших" элементов, и, как следствие, они имеют большие по модулю значения невязок. Следовательно, отбросив элементы выборки, соответствующие первым $m1$ и последним $m2$ элементам

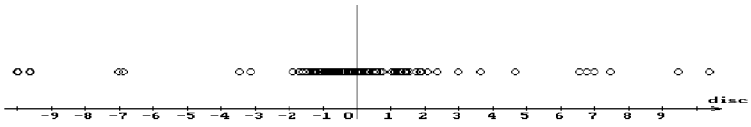

Рис. 1. Распределение $\epsilon_i, i = \overline{1,70}$

вариационного ряда (6), можно получить робастную оценку регрессии. Возникает вопрос, как определить $m1$ и $m2$, то есть определить выбросы в рабочей выборке. В отсутствии информации о допустимом отклонении оценок элементов V предлагается следующий подход к их дифференциации [2]. Применяя оценку Розенблата-Парзена [10], восстановим функцию плотности невязок $\epsilon_i, i = \overline{1,n}$

$$p_n(\epsilon, C_n^\epsilon) = \frac{1}{nC_n^\epsilon} \sum_{i=1}^{n} \Phi\Big(\frac{\epsilon - \epsilon_i}{C_n^\epsilon}\Big). \tag{7}$$

Исследование $p_n(\epsilon, C_n^\epsilon)$ на ближайший слева и справа к нулю минимумы (назовем их левым и правым соответственно), позволяет разделить обучающую выборку. Элементы, значения невязок которых лежат между левым и правым минимумами, составят "очищенную" выборку, остальные — выбросы.

Таким образом робастная оценка регрессии имеет вид

$$\tilde{y}_n(x, C_n, C_n^\epsilon) = \frac{\sum\limits_{i=1}^{n} y_i \prod\limits_{j=1}^{k} \Phi\big(\frac{x^j - x_i^j}{C_n}\big) I(\epsilon_i, C_n^\epsilon)}{\sum\limits_{i=1}^{n} \prod\limits_{j=1}^{k} \Phi\big(\frac{x^j - x_i^j}{C_n}\big) I(\epsilon_i, C_n^\epsilon)}, \tag{8}$$

$I(\cdot)$ — индикаторная функция, дифференцирующая элементы выборки,

$$I(\epsilon_i, C_n^\epsilon) = \begin{cases} 0; & \epsilon_i \in [a, c] \cup [d, b]; \\ 1; & \epsilon_i \in (c, d). \end{cases} \tag{9}$$

$a = \min\{\epsilon_i\}, i = \overline{1,n}; \quad b = \max\{\epsilon_i\}, i = \overline{1,n}; \quad c = \max\{\epsilon < 0\} : p_n(\epsilon, C_n^\epsilon) = \min \forall \epsilon < 0; \quad d = \min\{\epsilon > 0\} : p_n(\epsilon, C_n^\epsilon) = \min \forall \epsilon > 0.$

3. Оптимизация робастной оценки

Неизвестными в оценке (8) являются параметры размытости C_n и C_n^ϵ. Последний собственно и определяет качество робастизации оценки (8) или качество "очистки" выборки. Для нахождения оптимального значения C_n^ϵ предлагается следующий критерий

$$W_{p_n} = W_{\tilde{y}_n}^2 (C_n, C_n^\epsilon) \to \min_{C_n^\epsilon}, \tag{10}$$

где

$$W_{\tilde{y}_n}^2 (C_n, C_n^\epsilon) = \frac{1}{n} \sum_{i=1}^{n} \big(y_i - \tilde{y}_n(x_i, C_n, C_n^\epsilon)\big)^2; \tag{11}$$

C_n — оптимальный.

Таким образом, сначала при каждом фиксированном значении C_n^ϵ в ходе скользящего экзамена решается задача минимизации (11) по параметру C_n, и затем исследуется на минимум (10).

В основе критерия (10) лежит уже упоминавшееся свойство локальности непараметрической оценки регрессии, которая является чувствительной к качеству наблюдений функции в окрестности восстанавливаемой точки. Чем точнее измерения, тем точнее восстановленное значение функции в точке. С другой стороны, недостаток измерений так же сказывается отрицательным образом на качестве оценки в точках и, как следствие, на росте величины (11). Причем к изменению C_n^ϵ в (11) являются чувствительными только слагаемые, соответствующие качественным измерениям восстанавливаемой зависимости. Слагаемые соответствующие выбросам, в силу значительно меньшей доли последних в обучающей выборке, практически инвариантны к значению C_n^ϵ и следовательно не определяют изменения значений (11) с изменением C_n^ϵ.

При известном оптимальном значении C_n^ϵ нахождение оптимального параметра размытости C_n для "очищенной" выборки состоит в минимизации функционала

$$W^2_{\tilde{y}_n}(C_n) = \frac{1}{n}\sum_{i=1}^{n}\bigl(y_i - \tilde{y}_n(x_i, C_n, C_n^\epsilon)\bigr)^2 I(\epsilon_i, C_n^\epsilon) \to \min_{C_n}. \tag{12}$$

При отсутствии выбросов критерий (10) остается также работоспособным. В этом случае оптимальным является такой параметр C_n^ϵ, что в категорию выбросов не попадает ни один элемент. Это свойство предлагаемого подхода является довольно ценным, делая предложенный подход устойчивым к возможно ошибочной априорной информации о наличии или отсутствии выбросов в обучающей выборке.

4. Ремонт данных

Под ремонтом данных понимается идентификация и последующая замена грубых измерений (выбросов) значениями робастной модели. Для этого настраивается робастная оценка (8), то есть находятся оптимальные параметры размытости C_n^ϵ и C_n. Затем все выборочные значения, индикаторная функция $I(\cdot)$ которых принимает нулевое значение, заменяются их оценками (8). Для восстановления искомой зависимости по "отремонтированной" выборке можно использовать непараметрическую оценку регрессии (1), где в качестве параметра размытости C_n используется оптимальный для робастной оценки (8).

Заключение

Таким образом предложена непараметрическая робастная оценка регрессии, и алгоритм ремонта данных. Разработанный подход к оптимизации качества "очистки" выборки позволяет говорить об оптимальности последних. Кроме того можно утверждать, что критерий (10) претендует на универсальность и может быть применен для оптимизации любых робастных оценок типа цензурирования с той лишь разницей, что настраиваемым параметром будет не C_n^ϵ как в данном случае, а параметр соответствующей оценки, отвечающий за качество "очистки" выборки. К достоинствам даннного критерия можно отнести и то, что робастная оценка может быть применена для восстановления зависимостей по слабо зашумленным данным без ущерба для качества относительно обычной оценки.

Проведенное численное исследование предложенного подхода робастизации свидетельствует о высоком качестве получаемых оценок. Представлены результаты работы алгоритма (8), (10), (12) на выборке с выбросами (рис. 2) и без них (рис. 3).

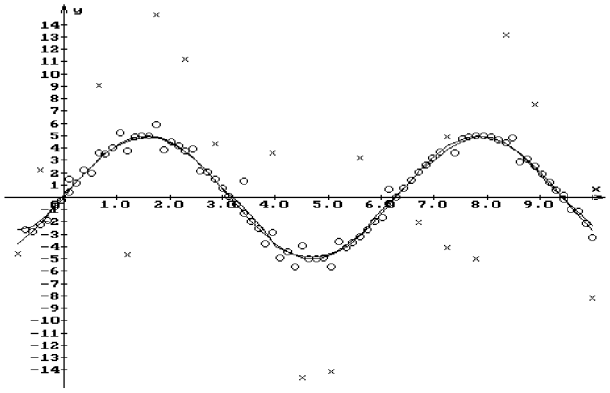

Рис. 2. Робастная оценка регрессии (8) и истинная зависимость, обучающая выборка содержит выбросы. (Согласно (10) крестики соответствуют выбросам и не не участвуют в вычислениях.)

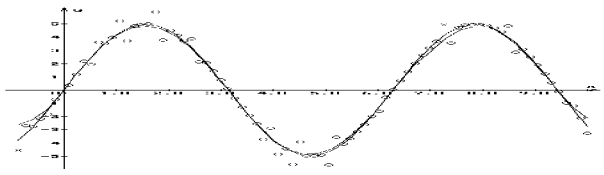

Рис. 3. Робастная оценка 8 и истинная зависимость, обучающая выборка без выбросов

Список литературы

[1] Ершов А. А. Стабильные методы оценки параметров. Обзор. Автоматика и телемеханика. 1978, № 5. С. 66–101.

[2] Кирик-Агапова Е. С. Об одном подходе к восстановлению и оптимизации робастных оценок функций //Тр. XXXII Региональной молодежной конф. Екатеринбург, ИММ УрО РАН. 2001. С. 31–37.

[3] Катковник В. Я. Непараметрическая идентификация и сглаживание данных. М.: Наука, 1985. 336 с.

[4] Медведев А. В. Непараметрические системы адаптации. Новосибирск: Наука, 1983. 174 с.

[5] Рубан А. И. Методы анализа данных: Учебное пособие. Ч. 1. Красноярск: КГТУ, 1994. 220 с.

[6] Смоляк С. А., Титаренко Б. П. Устойчивые методы оценивания. М.: Статистика, 1980.

[7] Box G. E. P. Non-normality and tests on variances. Biometrica, 1953, Vol. 40. P. 318–335.

[8] Gorban' A. N., Rossiev A. A. Neural network iterative method of principal curves for data with gaps // J. of Computer and System Sciences International. 1999. Vol. 38, No. 5. P. 825–850.

[9] Huber P. J. Robust statistics: a review. Ann. Math. Statistics, 1972. Vol. 43. P. 1041–1067.

[10] Parzen E. On estimation of probability density function and mode // Ann. Math. Stat. 1962. Vol. 33. P. 1065–1076.

[11] Rozenblatt M. Remarks on some nonparametric estimates of density function // Ann. Math. Stat. 1956. Vol. 27. P. 832–837.

[12] Rousseeuw P. J., van Zomeren B. C. Unmasking multivariate outliers and leverage points // J. of the American Statiscal Association. 1990. No. 85. P. 633–651.

[13] Tukey J. W. The future of the data analysis // Ann. Math. Stat. 1962. Vol. 33, No. 1. P. 1–67.