

КОЛЛЕКТИВ НЕПАРАМЕТРИЧЕСКИХ РЕГРЕССИЙ, ОСНОВАННЫЙ НА ДЕКОМПОЗИЦИИ ОБУЧАЮЩЕЙ ВЫБОРКИ

Предложена методика синтеза и анализа коллектива непараметрических регрессий, обеспечивающего высокую вычислительную эффективность решения задач восстановления стохастических зависимостей за счет использования технологии параллельных вычислений. Исследованы асимптотические свойства этого коллектива, приведены результаты их сравнения со свойствами традиционной непараметрической регрессии.

Ключевые слова: непараметрическая регрессия, большие выборки, асимптотические свойства, принципы декомпозиции, коллективное оценивание, параллельные вычислительные технологии.

Использование непараметрических моделей и алгоритмов, основанных на оценках плотности вероятности типа Розенблатта–Парзена, является одним из активно развивающихся направлений теории обучающихся систем. Эти модели и алгоритмы позволяют создавать типовые информационные средства, адаптируемые к условиям функционирования объектов различной природы. Однако при увеличении объема обучающей выборки вычислительная эффективность непараметрических статистик снижается. Подобные ситуации часто встречаются, например, при обработке больших массивов аэрокосмической информации. В этих условиях использование традиционных непараметрических моделей приводит к значительным временным затратам на формирование решений, в связи с чем возникает необходимость в разработке методики синтеза и анализа коллектива непараметрических регрессий, основанного на декомпозиции обучающих выборок по их объему.

Синтез коллектива непараметрических регрессий.

Пусть $V = (x^i, y^i, i = 1, n)$ – выборка, составленная из n независимых наблюдений случайных величин $x = (x_v, v = 1, k)$ и y , распределенных с неизвестными плотностями вероятности $p(x, y)$ и $p(x) > 0$. Вид однозначной стохастической зависимости

$$y = \varphi(x_1, \dots, x_k) \quad (1)$$

априори не задан.

Разобьем выборку V на T групп наблюдений $V_j = (x^i, y^i, i \in I_j)$, $j = 1, T$. Здесь I_j – множество номеров наблюдений переменных (x, y) , составляющих j -ю группу, причем $\bigcup_{j=1}^T I_j = I = (1, n)$.

На основании каждой выборки V_j осуществим синтез непараметрической регрессии [1]:

$$\bar{y}_j = \bar{\varphi}_j(x) = \frac{\sum_{i \in I_j} y^i \prod_{v=1}^k \Phi\left(\frac{x_v - x_v^i}{c_v(j)}\right)}{\sum_{i \in I_j} \prod_{v=1}^k \Phi\left(\frac{x_v - x_v^i}{c_v(j)}\right)}, \quad j = \overline{1, T}, \quad (2)$$

где $\Phi(u)$ – ядерные функции, удовлетворяющие условиям H :

$$\Phi(u) = \Phi(-u), \quad 0 \leq \Phi(u) < \infty,$$

$$\int \Phi(u) du = 1, \quad \int u^2 \Phi(u) du = 1,$$

$$\int u^m \Phi(u) du < \infty, \quad 0 \leq m < \infty;$$

$C_v(j)$, $v = \overline{1, k}$ – коэффициенты размытости ядерных функций, значения которых убывают с ростом количества элементов n_j множеств I_j , $j = \overline{1, T}$. Здесь и далее бесконечные пределы интегрирования опускаются.

Рассмотрим ряд модификаций статистической оценки зависимости (1):

– среднее значение $\bar{y}_j = \bar{\varphi}_j(x)$, $j = \overline{1, T}$. В качестве приближения $y = \varphi(x)$ по статистической выборке V используется статистика

$$\bar{y} = \bar{\varphi}(x) = \frac{1}{T} \sum_{j=1}^T \bar{\varphi}_j(x). \quad (3)$$

Оптимизация частных непараметрических регрессий (2) по коэффициентам размытости c_v , $v = \overline{1, k}$, ядерных функций осуществляется в режиме «скользящего экзамена» исходя из условия минимума статистической оценки точности аппроксимации зависимости (1)

$$\bar{W}_j = \frac{1}{n_j} \sum_{i \in I_j} (y^i - \bar{\varphi}_j(x^i))^2.$$

При формировании слагаемых $\bar{W}_j$ ситуация (x^i, y^i) исключается из процесса обучения в выражении непараметрической регрессии (2);

– средневзвешенное значение $\bar{y}_j$, $j = \overline{1, T}$. В этом случае обобщенная модель искомой зависимости формируется в соответствии с выражением

$$\tilde{y} = \sum_{j=1}^T \alpha_j \bar{\varphi}_j(x),$$

где $\alpha_j = \frac{\bar{W}^{-1}}{\sum_{i=1}^T \bar{W}_i^{-1}}$ – весовые коэффициенты частных непараметрических регрессий $\bar{\varphi}_j(x)$, $j = \overline{1, T}$, сумма которых равна единице;

– оценивание наиболее вероятного значения искомой зависимости. Будем полагать, что значение T достаточно для оценивания по выборке y_{j2} , $j = \overline{1, T}$, плотности вероятности частных решений $p(y)$ при конкретном значении x . Для этого используем непараметрическую оценку плотности вероятности типа Розенблатта–Парзена [2]:

$$\bar{p}(y) = \frac{1}{Tc} \sum_{j=1}^T \Phi\left(\frac{y - \bar{y}_j}{c}\right). \quad (4)$$

Выбор оптимального параметра c статистики (4) осуществляется по условию максимума функции правдоподобия

$$L(c) = \prod_{i=1}^T \bar{p}(\bar{y}^{-i}).$$

При вычислении значений $\bar{p}(\bar{y}^{-i})$ в соответствии со статистикой (4) полагаем $i \neq j$.

Наиболее вероятное значение y находим по условию

$$y = \arg \max_{\bar{y}} \bar{p}(\bar{y}),$$

используя численные методы поиска экстремума – максимума оценки плотности вероятности $\bar{p}(\bar{y})$.

Асимптотические свойства коллектива непараметрических регрессий. Без существенной потери общности рассмотрим асимптотические свойства статистики (3) для случая, когда x – скаляр и вид плотности вероятности $p(x)$ известен. В этих условиях справедливо следующее утверждение.

Теорема. Пусть $\varphi(x), p(x, y), p(x) \neq 0$ и первые их производные ограничены и непрерывны; ядерные функции $\Phi(u) \in H$; последовательности $c_j = c(n_j)$ коэффициентов размытости ядерных функций таковы, что при $n_j \rightarrow \infty$ значения $c(n_j) \rightarrow 0$, а значения $n_j c_j \rightarrow \infty, j = 1, T$. Тогда при конечных значениях T коллектив непараметрических регрессий (3) обладает свойством асимптотической несмещенности и состоятельности.

Доказательство. По определению имеем

$$\begin{aligned} M(\bar{\varphi}(x)) &= \frac{1}{T} \sum_{j=1}^T M(\bar{\varphi}_j(x)) = \\ &= \frac{1}{T} \sum_{j=1}^T (c_j^{-1} p(x))^{-1} \int \varphi(t) \Phi\left(\frac{x-t}{c_j}\right) p(t) dt, \end{aligned}$$

где $\varphi(t) = M\left(\frac{y}{x}\right)$; M – знак математического ожидания.

Проведем в интеграле замену переменных $(x-t)c_j^{-1} = u$ и, разлагая функции $\varphi(x - cu), p(x - cu)$ в ряд Тейлора в точке x , с учетом свойств ядерной функции $\Phi(u)$ получим

$$\begin{aligned} W_1 &= M(\bar{\varphi}(x) - \varphi(x)) \sim A^1(x) \times \\ &\times \frac{1}{T} \sum_{j=1}^T c_j^2 + A^2(x) \frac{1}{T} \sum_{j=1}^T c_j^4. \end{aligned} \quad (5)$$

Здесь

$$\begin{aligned} A^1(x) &= (\varphi(x) p(x))^{(2)} / (2p(x)), \\ A^2(x) &= p^{(2)}(x) \varphi^{(2)}(x) \int u^4 \Phi(u) du / (4p(x)), \end{aligned}$$

где $p^{(2)}(x), \varphi^{(2)}(x)$ – вторые производные по x соответствующих функций.

Отсюда при выполнении условий $c(n_j) \rightarrow 0$ и $n_j \rightarrow \infty$ следует свойство асимптотической несмещенности статистики (3).

Для доказательства состоятельности оценки $\bar{y} = \bar{\varphi}(x)$ вычислим ее дисперсию:

$$D(\bar{\varphi}(x)) = M(\bar{\varphi}(x) - \varphi(x))^2 - (M(\bar{\varphi}(x) - \varphi(x)))^2.$$

Исследуем асимптотические свойства среднеекватического отклонения:

$$\begin{aligned} M(\bar{\varphi}(x) - \varphi(x))^2 &= \\ &= \frac{1}{T^2} M \left(\sum_{j=1}^T (\varphi(x) - \bar{\varphi}_j(x))^2 + \right. \\ &\quad \left. + \sum_{j=1}^T \sum_{t=1}^T (\varphi(x) - \bar{\varphi}_j(x)) (\varphi(x) - \bar{\varphi}_t(x)) \right). \end{aligned} \quad (6)$$

С учетом результатов исследования асимптотических свойств традиционной непараметрической регрессии, пренебрегая величинами малости $c_j/n_j, c_j^2 c_t^4, j = 1, T, t = 1, T$, приведем выражение (6) при достаточно больших значениях $n_j, j = 1, T$, к виду

$$\begin{aligned} M(\bar{\varphi}(x) - \varphi(x))^2 &= \\ &= \frac{1}{T^2} \left(\sum_{j=1}^T \left(\frac{\|\Phi(u)\|^2 \varphi^2(x)}{n_j c_j p(x)} + c_j^4 (A^1(x))^2 \right) + \right. \\ &\quad \left. + (A^1(x))^2 \sum_{j=1}^T \sum_{t=1}^T c_j^2 c_t^2 \right). \end{aligned} \quad (7)$$

Нетрудно заметить, что при выполнении условий $c(n_j) \rightarrow 0$ и $n_j c_j \rightarrow \infty$ при $n_j \rightarrow \infty, j = 1, T$, коллектив непараметрических регрессий (3) сходится в среднеекватическом отклонении, а с учетом свойств асимптотической несмещенности является состоятельным.

Сравнение аппроксимационных свойств коллектива непараметрических регрессий и традиционной непараметрической регрессии. Рассмотрим соотношения между асимптотическими смещениями, среднеекватическими отклонениями и дисперсиями непараметрической регрессии $\tilde{y} = \tilde{\varphi}(x)$ и $y = \varphi(x)$ при оптимальных значениях коэффициента размытости ядерных функций.

Асимптотическое выражение смещения для непараметрической регрессии при $x \in R^1$

$$\tilde{y} - \varphi(x) = (ncp(x))^{-1} \sum_{i=1}^n y^i \Phi\left(\frac{x-x^i}{c}\right) \quad (8)$$

имеет вид [3]

$$W_2 = M(\tilde{\varphi}(x) - \varphi(x)) \sim c^2 A^1(x) + c^4 A^2(x),$$

а для обобщенной непараметрической регрессии (3) определяется выражением (5).

Оптимальное значение c^* статистики (8), минимизирующее асимптотическое выражение среднеекватического отклонения

$$\int M(\tilde{\varphi}(x) - \varphi(x))^2 dx, \quad (9)$$

определяется по формуле

$$c^* = \left(\frac{\|\Phi(u)\|^2 \int \varphi^2(x) p^{-1}(x) dx}{n \int (A^1(x))^2 dx} \right)^{1/5},$$

где $\|\Phi(u)\|^2 = \int \Phi^2(u) du$.

Для получения аналитически значимых результатов примем $n_j = n/T, j = 1, T$. В этом случае оптимальные значения $\bar{c}(n_j) = \bar{c}(n/T) = \bar{c}$ частных непараметрических регрессий $\bar{\varphi}_j(x), j = 1, T$, в смысле минимума критерия типа (9) имеют вид $\bar{c} = c^* T^{1/5}$.

Нетрудно показать, что при оптимальных значениях коэффициентов размытости разность между W_1 и W_2 дана выражением

$$\begin{aligned} \Delta W_{21} &= W_2 - W_1 = (c^*)^2 (1 - T^{2/5}) \times \\ &\times [A^1(x) + A^2(x) (c^*)^2 (1 + T^{2/5})]. \end{aligned}$$

Знак ΔW_{21} определяется разностью $(1 - T^{2/5}) < 0$ при $T > 1$, т. е. смещение обобщенной непараметрической регрессии $\varphi(x)$ больше, чем $\tilde{\varphi}(x)$.

Как и следовало ожидать, при $T=1$ асимптотические смещения $\bar{\varphi}(x)$ и $\tilde{\varphi}(x)$ одинаковы.

При разбиении исходной выборки на T групп точек одинакового объема n/T асимптотическое выражение критерия

$$W_3 = \int M \left(\bar{\varphi}(x) - \varphi(x) \right)^2 dx$$

при оптимальных коэффициентах $\bar{c}^*$ ядерных функций преобразуется к виду

$$W_3 \sim \left(\frac{\left(\frac{\| \Phi(u) \|^2 \int \varphi^2(x) p^{-1}(x) dx}{n} \right)^4}{\int (A^1(x))^2 dx} \right)^{1/5} \times \times \frac{1+T^2}{T^{6/5}} = W_4 \frac{1+T^2}{T^{6/5}}. \quad (10)$$

При этом W_3 связано с асимптотическим выражением W_4 критерия (9) для традиционной непараметрической регрессии соотношением

$$W_3 = \frac{1+T^2}{2T^{6/5}} W_4.$$

Тогда

$$\frac{W_4}{W_3} = \frac{2T^{6/5}}{1+T^2} < 1, \quad (11)$$

что подтверждает несколько большую эффективность в среднеквадратическом отклонении непараметрической регрессии (8) по сравнению с коллективом (3), например при $T=2$ отношение (11) равно 0,92, а при $T=3-0,75$.

Вместе с тем дисперсия статистики (8) больше, чем дисперсия коллектива непараметрических регрессий (3).

В этом нетрудно убедиться, если сравнить их главные дисперсионные составляющие в асимптотическом выражении критерия (9) при оптимальных коэффициентах размытости в W_6 и значение $W_5 = \frac{W_6}{T^{6/5}}$ в соответствующем выражении критерия (7). Их отношение

$$\frac{W_6}{W_5} = T^{6/5} > 1, \quad \forall T > 1.$$

Как и следовало ожидать, коллектив непараметрических регрессий (3) характеризуется меньшей дисперсией по сравнению со статистикой (8).

Коллектив непараметрических регрессий, основанный на декомпозиции обучающей выборки по объему, обеспечивает возможность использования технологии параллельных вычислений при восстановлении многомерных стохастических зависимостей. Установлено, что этот коллектив имеет асимптотические свойства, а его дисперсия значительно меньше дисперсии традиционной непараметрической регрессии.

Библиографический список

1. Надарая, Э.А. Непараметрические оценки кривой регрессии / Э.А. Надарая // Тр. ВЦ АН ГССР, 1965. Вып. 5. С. 56–68.
2. Parzen, E. On estimation of a probability density function and mode / E. Parzen // Ann. Math. Statistic. 1962. Vol. 33. P. 1065–1076.
3. Лапко, А. В. Имитационные модели неопределенных систем / А.В. Лапко. Новосибирск: Наука, Сиб. отд-ние, 1993.

A. V. Lapko, V. A. Lapko, S. S. Varochkin

COLLECTIVE OF NONPARAMETRIC REGRESSIONS BASED ON DECOMPOSITION OF TRAINING SAMPLE

The technique of synthesis and the analysis of collective nonparametric regresses providing high computing efficiency of the decision of problems of restoration of stochastic dependences due to use of technology of parallel calculations is offered. Asymptotic properties of collective are researched, its comparison with traditional nonparametric regression is made.

Keywords: nonparametric regression, the big samples, asymptotic properties, principles of decomposition, collective estimations, parallel computing technologies.