

УДК 519.7

А.В. Лапко, В.А. Лапко

ОБОБЩЁННАЯ НЕПАРАМЕТРИЧЕСКАЯ РЕГРЕССИЯ И ЕЁ СВОЙСТВА

Рассматривается методика синтеза и анализа обобщённой непараметрической регрессии. Идея рассматриваемого подхода состоит в построении упрощённых параметрических моделей относительно некоторого набора точек из обучающей выборки с последующей их интеграцией на основе методов непараметрической статистики. Исследуются свойства полученной аппроксимации.

Ключевые слова: непараметрическая регрессия, коллективное оценивание, упрощённые аппроксимации, асимптотические свойства

На современном этапе развития теории обучающихся систем настойчиво обсуждается и разрабатывается идея о совместном использовании разнотипных моделей – как средства наиболее полного учета априорной информации. Известно яркое высказывание профессора В. Хардле [1]: «Совмещение параметрических и непараметрических составляющих может даже привести к построению лучшей модели, чем непараметрический или параметрический подход!». Получены первые успешные результаты исследований в данном направлении, к которым можно отнести методы локальной аппроксимации [2], гибридные модели [3–5], полупараметрические и частично линейные модели [1].

Разработан новый класс непараметрических моделей коллективного типа для решения задач восстановления стохастических зависимостей [6], распознавания образов [7] и анализа временных процессов [8]. Синтез подобных моделей сводится к непараметрическому оцениванию функционалов от семейства регрессий, построенных относительно системы «опорных» точек из экспериментальных данных. Их применение позволяет в наиболее полном объёме использовать информацию обучающей выборки, содержащуюся в её элементах, и взаимосвязи между ними.

Цель данной работы стоит в развитии методики синтеза обобщённой непараметрической регрессии, основанной на сочетании преимуществ параметрических и локальных аппроксимаций восстанавливаемой функции и исследовании её свойств.

1. Синтез обобщённой непараметрической регрессии

Пусть дана выборка $V = (x^i, y^i, i = \overline{1, n})$ из статистически независимых наблюдений значений y^i неизвестной однозначной зависимости

$$y = F(x) \quad \forall x \in R^k \quad (1)$$

и её аргументов x^i .

Полагается, что элементы выборки V проверены на наличие ошибок контроля и последние удалены. Причём соблюдается условие $y^i \neq 0, i = \overline{1, n}$, необходимое

при использовании относительной ошибки аппроксимации восстанавливаемой зависимости.

Поставим в соответствие некоторым точкам (x^i, y^i) обучающей выборки V , условно назовём их опорными, упрощённые аппроксимации $\varphi_i(x, \bar{\alpha})$ зависимости (1), параметры α которых удовлетворяют условиям

$$y^i = \varphi_i(x^i, \bar{\alpha}^i),$$

$$\bar{\alpha}^i = \arg \min_{\alpha^i} \frac{1}{n-1} \sum_{\substack{j=1 \\ j \neq i}}^n (y^j - \varphi_i(x^j, \alpha^i))^2, \quad i = \overline{1, N}. \quad (2)$$

Упрощённые аппроксимации $\varphi_i(x, \bar{\alpha}^i)$, например линейные, проходят через опорные точки $(x^i, y^i, i = \overline{1, N})$ и близки в среднеквадратическом к остальным элементам обучающей выборки V .

Примем в качестве статистической модели зависимости (1) процедуру условного усреднения

$$\bar{\varphi}(x) = \sum_{i=1}^N \varphi_i(x, \bar{\alpha}^i) \lambda^i(x), \quad (3)$$

где положительная функция $\lambda^i(x)$ определяет «вес» правила $\varphi_i(x, \bar{\alpha}^i)$ при формировании решения в ситуации x . Причём сумма $\lambda^i(x)$, $i = \overline{1, N}$, равна единице.

Примером функции $\lambda^i(x)$ является «весовая» функция

$$\lambda^i(x) = \frac{\prod_{v=1}^k \Phi\left(\frac{x_v - x_v^i}{c_v}\right)}{\sum_{i=1}^N \prod_{v=1}^k \Phi\left(\frac{x_v - x_v^i}{c_v}\right)},$$

составленная из положительных, нормированных и симметричных «ядерных» функций $c_v^{-1} \Phi\left(\frac{x_v - x_v^i}{c_v}\right)$, на основе которых строятся непараметрические модели [9].

В этом случае при $\varphi_i(x, \bar{\alpha}^i) = y^i$ и $N = n$ статистика (3) преобразуется в традиционную непараметрическую регрессию

$$\tilde{y} = \tilde{\varphi}(x) = \sum_{i=1}^n y^i \lambda^i(x).$$

Проведём анализ обобщённой непараметрической регрессии (3) для линейных упрощённых аппроксимаций

$$\varphi_i(x, \alpha^i) = \sum_{v=1}^k \alpha_v^i x_v + \beta_i, \quad i = \overline{1, N}.$$

Заметим, что в соответствии с методикой синтеза статистики $\bar{\varphi}(x)$ (3)

$$\beta_i = y^i - \sum_{v=1}^k \alpha_v^i x_v^i,$$

а параметры $\bar{\alpha}_v^i$, $i = \overline{1, N}$, определяются из условия минимума критерия (2). Тогда

$$\varphi_i(x, \bar{\alpha}^i) = y^i + \sum_{v=1}^k \bar{\alpha}_v^i (x_v - x_v^i), \quad i = \overline{1, N}. \quad (4)$$

Подставляя упрощённые аппроксимации (4) в выражение (3), получим

$$\bar{\varphi}(x) = \sum_{i=1}^N y^i \lambda^i(x) + \sum_{i=1}^N \sum_{v=1}^k \bar{\alpha}_v^i (x_v - x_v^i) \lambda^i(x). \quad (5)$$

Первое слагаемое в выражении (5) представляет собой непараметрическую регрессию, обладающую свойствами асимптотической сходимости к условному математическому ожиданию – оптимальной модели (1) в среднеквадратическом смысле [10, 11].

Вторая составляющая (5) играет роль поправочного члена и отражает условную взаимосвязь между точками обучающей выборки. Его значения, в соответствии с особенностями первого слагаемого (5), снижаются по мере роста объёма исходной информации, что подтверждается результатами аналитических исследований [11]. Наличие поправочного члена делает статистику (3) схожей с гибридными моделями, а слабая зависимость её свойств от вида опорных функций – с непараметрической регрессией.

Для синтеза обобщённой непараметрической регрессии разработана итерационная процедура формирования системы опорных точек. Идея предлагаемого метода основывается на последовательном анализе относительных расхождений между значениями восстанавливаемой зависимости и строящейся обобщённой непараметрической регрессии $\Psi_t(\varphi_j(x), j = \overline{1, t})$:

$$W(i) = \left| \frac{y^i - \Psi_t(\varphi_j(x^i), j = \overline{1, t})}{y^i} \right|, \quad i \in I_{\bar{t}},$$

где $I_{\bar{t}} = I \setminus I_t$ – множество номеров точек, не входящих в число опорных с номерами из множества I_t , а I – множество номеров точек исходной выборки. Если модель $\Psi_t(\varphi_j(x^i), j = \overline{1, t})$ в некоторой точке x^i имеет максимальное расхождение с экспериментальным значением y^i , то естественно принять эту точку (x^i, y^i) в качестве опорной при построении $(t+1)$ упрощённой аппроксимации.

Процедура формирования опорных точек заканчивается на t -й итерации, когда ошибка аппроксимации

$$\bar{W}(\Psi_t(\varphi_j(x), j = \overline{1, t})) = \frac{1}{|I_{\bar{t}}|} \sum_{i \in I_{\bar{t}}} W(i) < \bar{\bar{W}}$$

меньше заданного порога $\bar{\bar{W}}$, удовлетворяющего пользователя. Здесь $|I_{\bar{t}}|$ – количество элементов множества $I_{\bar{t}}$.

2. Модификация обобщённой непараметрической регрессии

Для повышения аппроксимационных свойств обобщённой непараметрической регрессии предлагается учитывать статистические оценки эффективности $\bar{W}^i$ упрощённых параметрических аппроксимаций $\varphi_i(x, \bar{\alpha}^i)$, $i \in I_o$. Здесь I_o – множество номеров опорных точек из выборки V . В качестве показателя эффективности i -й аппроксимации может выступать среднеквадратический критерий

$$\bar{W}^i = \frac{1}{n-1} \sum_{\substack{j=1 \\ j \neq i}}^n \left(y^j - \varphi_i(x^j, \bar{\alpha}^i) \right)^2, \quad i \in I_o.$$

Учёт эффективности целесообразно осуществить, вводя в «весовую» функцию $\lambda^i(x)$ ядерную меру близости между значением $\bar{W}^i$ и её минимальным значением (нулём). В результате полученная модификация обобщённой непараметрической регрессии (3) с учётом оценок эффективности упрощённых параметрических аппроксимаций имеет вид

$$\bar{\varphi}(x) = \frac{\sum_{i \in I_o} \varphi_i(x, \bar{\alpha}^i) \prod_{v=1}^k \Phi\left(\frac{x_v - x_v^i}{c_v}\right) \Phi\left(\frac{0 - \bar{W}^i}{c_w}\right)}{\sum_{i \in I_o} \prod_{v=1}^k \Phi\left(\frac{x_v - x_v^i}{c_v}\right) \Phi\left(\frac{0 - \bar{W}^i}{c_w}\right)},$$

где c_w – параметр ядерной функции $\Phi\left(\frac{0 - \bar{W}^i}{c_w}\right)$, который характеризует область её определения.

3. Асимптотические свойства обобщённой непараметрической регрессии

Для удобства последующего анализа предположим, что x – скаляр и закон распределения $p(x)$ известен. Тогда обобщённая непараметрическая регрессия типа (3) будет иметь вид

$$\bar{\varphi}(x) = \frac{1}{Ncp(x)} \sum_{i=1}^N \varphi_i(x, \bar{\alpha}^i) \Phi\left(\frac{x - x^i}{c}\right). \quad (6)$$

Если принять

$$\varphi_i(x, \bar{\alpha}) = \bar{\alpha}^i x + \beta^i,$$

то при выполнении условия (2) её прохождения через опорную точку

$$\beta^i = y^i - \bar{\alpha}^i x^i.$$

При этом упрощённая аппроксимация $\varphi_i(x, \bar{\alpha})$ соответствует выражению

$$\varphi_i(x, \bar{\alpha}) = y^i + \bar{\alpha}^i (x - x^i).$$

Тогда модель (6) запишется в виде статистики

$$\bar{\varphi}(x) = \frac{1}{Ncp(x)} \sum_{i=1}^N y^i \Phi\left(\frac{x - x^i}{c}\right) + \frac{1}{Ncp(x)} \sum_{i=1}^N \bar{\alpha}^i (x - x^i) \Phi\left(\frac{x - x^i}{c}\right). \quad (7)$$

Теорема. Пусть 1) кривая регрессии $\varphi(x)$ и плотности вероятностей $p(x, y)$, $p(x)$, характеризующие распределения переменных x , y исходных статистических данных и опорных точек обобщённой непараметрической регрессии, являются ограниченными и непрерывными со всеми своими производными до второго порядка включительно; 2) ядерные функции $\Phi(u)$ являются положительными, симметричными и нормированными; 3) объём исходных статистических данных $n \rightarrow \infty$; 4) последовательность $c = c(N) \rightarrow 0$ при $N \rightarrow \infty$, а $Nc \rightarrow \infty$.

Тогда:

смещение

$$M(\bar{\varphi}(x) - \varphi(x)) \sim c^2 \frac{A_1(x, y) + A(x, y)}{2p(x)D(x)}; \quad (8)$$

квадратическое отклонение

$$M(\bar{\varphi}(x) - \varphi(x))^2 \sim \frac{\varphi^2(x) \|\Phi(u)\|^2}{Ncp(x)} + \frac{c^4}{p(x)} \left[\frac{((\varphi(x)p(x))^{(2)})^2}{4p(x)} + \frac{A(x, y)}{D(x)} \times \right. \\ \left. \times \left(\frac{A(x, y)}{4p(x)D(x)} + A_1(x, y) \right) \right], \quad (9)$$

где M – знак математического ожидания; $A(x, y)$, $A_1(x, y)$ – нелинейные функционалы от $\varphi(x)$, $p(x, y)$, $p(x)$ и их производных; $D(x)$ – дисперсия опорных точек; $\|\Phi(u)\|^2 = \int \Phi^2(u) du$. Здесь и далее бесконечные пределы интегрирования опускаются.

Из выражений (8), (9) при выполнении условий теоремы следует асимптотическая несмещённость, сходимость в среднеквадратическом и состоятельность обобщённой непараметрической регрессии.

Для доказательства данных утверждений используется технология преобразований, предложенная В.А. Епанечниковым при исследовании асимптотических свойств непараметрической оценки плотности вероятности [12] и развитая в работах [13–15].

В исследуемую модель (7) подставляются оптимальные значения параметров

$$\bar{\alpha}^i = \sum_{\substack{j=1 \\ j \neq i}}^n (y^i - y^j)(x^i - x^j) \bigg/ \sum_{\substack{j=1 \\ j \neq i}}^n (x^i - x^j)^2, \quad i = \overline{1, N},$$

минимизирующих критерий (2).

При выполнении последующих преобразований учитывается, что элементы исходной статистической выборки V и множества опорных точек являются значениями одних и тех же случайных величин (t, y) с плотностью вероятности $p(t, y)$.

Определим кривую регрессии

$$\varphi(t) = \int y p(y/t) dy$$

и проведём замену переменных $(x - t)c^{-1} = u$ в составляющих $M(\bar{\varphi}(x) - \varphi(x))$ и $M(\bar{\varphi}(x) - \varphi(x))^2$.

Разложим функции $\varphi(x - ci)$, $p(x - ci)$ в ряды Тейлора в точке x и произведём интегрирование полученных выражений с учётом свойств ядерных функций.

Далее, проводя достаточно громоздкие вычисления и отбросив составляющие высших порядков малости при $n \rightarrow \infty$, $N \rightarrow \infty$ получим приведённые асимптотические выражения смещения (8) и квадратического отклонения (9).

Установлено, что асимптотические свойства обобщённой непараметрической регрессии несущественно зависят от объёма выборки, используемой при идентификации упрощённых аппроксимаций. Эффективность предлагаемой модели в основном определяется законом распределения системы опорных точек и их количеством.

Сравнение аппроксимационных свойств исследуемой модели $\bar{y} = \bar{\varphi}(x)$ и традиционной непараметрической регрессии $\tilde{y} = \tilde{\varphi}(x)$ осуществляется путём анализа отношения соответствующих им среднеквадратических критериев

$$R = \frac{M \int (\varphi(x) - \tilde{\varphi}(x))^2 dx}{M \int (\varphi(x) - \bar{\varphi}(x))^2 dx}$$

при $n \rightarrow \infty$ и $N \rightarrow \infty$.

При выполнении условия $R > 1$ имеет место преимущество $\bar{\varphi}(x)$ над $\tilde{\varphi}(x)$. Поэтому ставится задача поиска условий синтеза обобщённой непараметрической регрессии, обеспечивающих выполнение данного требования.

Используемая схема исследований предполагает выполнение следующих действий.

Проинтегрируем выражение (9) и обозначим его через $\bar{W}$. Определим оптимальное значение коэффициентов размытости $\bar{c}$, $\tilde{c}$ ядерных функций статистик $\bar{\varphi}(x)$ и $\tilde{\varphi}(x)$, минимизирующих соответственно $\bar{W}$ и асимптотическое выражение среднеквадратического критерия [15]

$$\tilde{W} = \frac{\|\Phi(u)\|^2}{nc} \int \varphi^2(x) p^{-1}(x) dx + \frac{c^4}{4} \int \left(p^{-1}(x) (\varphi(x) p(x))^{(2)} \right)^2 dx.$$

Вычислим минимальное значение $\bar{W}_0$, $\tilde{W}_0$ выражений $\bar{W}$, $\tilde{W}$ при оптимальных $\bar{c}$, $\tilde{c}$. Проведём анализ отношения $\bar{R} = \frac{\tilde{W}_0}{\bar{W}_0}$ и установим условия, при которых $\bar{R} > 1$.

Для равномерных законов распределения аргументов восстанавливаемой зависимости и опорных точек это условие определяется неравенством

$$\left(\frac{N}{n} \right)^4 > 11 \left(\frac{y_{\max}^{(1)}}{D(x)} \right)^2 + 7 \left(\frac{y_{\max}}{D(x)} \right)^2 + 4 \frac{y_{\max}^{(1)}}{D(x)} + \frac{2}{3} \frac{y_{\max}^{(1)} y_{\max}}{D^2(x)},$$

где $y_{\max}$, $y_{\max}^{(1)}$ – максимальные значения восстанавливаемой зависимости и её производной.

4. Анализ результатов вычислительных экспериментов

Проведём сравнение аппроксимационных свойств традиционной и обобщённой непараметрических регрессий при восстановлении стохастической зависимости

$$F(x) = 1 - x + \exp(-20(x - 0,5)^2),$$

приведённой в работе [1].

При формировании исходных данных $V = (x^i, y^i, i = \overline{1, n})$ объёма n значения $x \in [0; 1]$ определялись с равномерным законом распределения, а на значения функции накладывалась помеха в соответствии с выражением

$$y^i = F(x^i) + 2(0,5 - \varepsilon^i)F(x^i)r, i = \overline{1, n}.$$

Здесь $r \in [0; 1]$ – уровень помех; $\varepsilon^i \in [0; 1]$ – случайная величина с равномерным законом распределения.

В статистике (3) при $k=1$ упрощённые аппроксимации принимались линейными, а функции $\Phi(u)$ соответствовали оптимальным в среднеквадратическом смысле ядрам В.А. Епанечникова [12]. В качестве показателя эффективности сравниваемых моделей использовалась среднеквадратическая ошибка типа

$$\tilde{W} = \frac{1}{\bar{n}} \sum_{j=1}^{\bar{n}} (F(x^j) - \bar{\Phi}(x^j))^2,$$

которая вычислялась по контрольной выборке объёма $\bar{n} = 1000$. При её оценивании для каждого условия эксперимента проводилось 50 имитаций обучающих выборок, а полученные результаты усреднялись.

При формировании обобщённой регрессии (3) опорные точки выбирались из исходной выборки V с равномерным законом распределения.

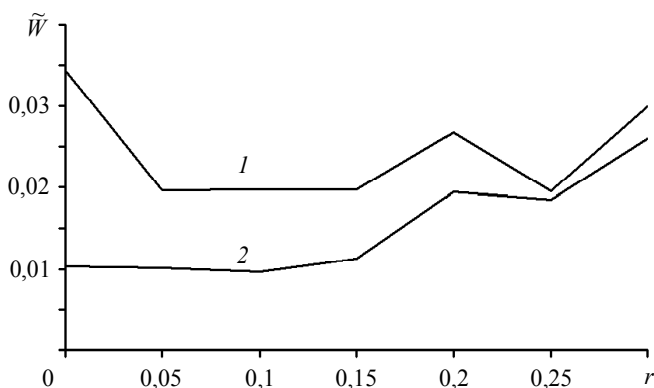


Рис. 1 Зависимость среднеквадратической ошибки $\tilde{W}$ от уровня помех r при объёме исходных данных $n=20$ и отношении $N/n=0,5$. Кр. 1, 2 соответствуют традиционной непараметрической регрессии и исследуемой статистике (3)

Установлено, что точность аппроксимации обобщённой непараметрической регрессии в основном зависит от количества опорных точек, закона их распределения и уровня помех r . Исследуемая статистика обладает преимуществом над традиционной непараметрической регрессией при относительно малых объёмах исходных данных ($N \leq 100$ в условиях $N/n \in (0,2; 0,5)$ и $r < 0,2$). При больших значениях n эффективность сравниваемых моделей сопоставима, что согласуется с результатами аналитических исследований. С ростом уровня помех качество аппроксимации исследуемой статистики снижается. Вместе с тем двойное сглаживание, используемое при синтезе модели (3), обеспечивает её преимущество над непараметрической регрессией, что особенно проявляется при малых значениях уровня помех (рис. 1).

Заключение

Обобщённая непараметрическая регрессия для решения задач восстановления стохастических зависимостей занимает промежуточное положение между локальными и параметрическими методами аппроксимации функций и использует их преимущества. Структуру изучаемого класса моделей составляют семейство упрощённых параметрических аппроксимаций исследуемой функции, каждая из которых строится относительно системы опорных ситуаций из обучающей выборки. Объединение упрощённых аппроксимаций осуществляется с помощью непараметрической оценки оператора условного математического ожидания.

Установлено, что асимптотические свойства обобщённой непараметрической регрессии слабо зависят от вида упрощённых аппроксимаций и объёма выборки в задаче их идентификации. Эффективность предлагаемых моделей в значительной степени определяется законом распределения системы опорных точек и их количеством.

Исследуемая статистика обладает преимуществом над традиционной непараметрической регрессией при относительно малых объёмах исходных данных. При больших значениях n эффективность сравниваемых моделей сопоставима, что согласуется с результатами аналитических исследований.

ЛИТЕРАТУРА

1. Хардле В. Прикладная непараметрическая регрессия. М.: Мир, 1993. 349 с.
2. Катковник В.Я. Линейные и нелинейные методы непараметрического регрессионного анализа // Автоматика. 1979. № 5. С. 165–170.
3. Лапко А.В., Лапко В.А., Ярославцев С.Г. Разработка и исследование гибридных алгоритмов в задачах распознавания образов // Автометрия. 2006. № 1. С. 32–39.
4. Лапко В.А., Лапко В.А. Гибридные модели стохастических зависимостей // Автометрия. 2002. № 5. С. 38–48.
5. Лапко А.В., Лапко В.А., Саренков А.В. Синтез и анализ линейных гибридных решающих функций в задаче распознавания образов // Системы управления и информационные технологии. 2012. № 1 (47). С. 66–69.
6. Лапко В.А. Синтез и анализ непараметрических моделей коллективного типа // Автометрия. 2001. № 6. С. 98–106.
7. Lapko V. A. Nonparametric models of pattern recognition of collective type // Pattern Recognition and Image Analysis. 2002. V. 12. No. 4. P. 354–361.
8. Лапко В.А. Непараметрические модели временных зависимостей, основанные на методе двойного коллективного оценивания // Автометрия. 2002. № 1. С. 42–50.

9. Parzen E. On estimation of a probability density function and mode // Ann. Math. Statistic. 1962. V. 33. No. 3. P. 1065–1076.
10. Нада́ря Э.А. Замечания о непараметрических оценках плотности вероятности и кривой регрессии // Теория вероятностей и её применение. 1970. Т. 15. № 1. С. 139–142.
11. Лапко А.В., Лапко В.А. Анализ асимптотических свойств многомерной непараметрической регрессии // Вестник СибГАУ. 2012. № 2 (42). С. 41–44.
12. Епанечников В.А. Непараметрическая оценка многомерной плотности вероятности // Теория вероятности и ее применения. 1969. Т. 14. № 1. С. 156–161.
13. Лапко А.В., Лапко В.А. Анализ асимптотических свойств непараметрической оценки уравнения разделяющей поверхности в двувальтернативной задаче распознавания образов // Автометрия. 2010. Т. 46. № 3. С. 48 – 53.
14. Лапко А.В., Лапко В.А. Анализ непараметрических алгоритмов распознавания образов в условиях пропуска данных // Автометрия. 2008. № 3. С. 65–74.
15. Лапко А.В., Лапко В.А. Коллектив многомерных непараметрических регрессий, основанный на декомпозиции обучающей выборки по её объёму // Вестник СибГАУ. 2012. 3 (43). С. 42–46.

Лапко Александр Васильевич

Лапко Василий Александрович

Институт вычислительного моделирования СО РАН,

Сибирский государственный аэрокосмический

университет им. академика М.Ф. Решетнева

E-mail: lapko@icm.krasn.ru

Поступила в редакцию 5 марта 2013 г.

Lapko Alexandr V., Lapko Vasily A. (Institute of Computational Modeling, Siberian Branch of Russian Academy of Sciences; Reshetnev Siberian State Aerospace University). **The generalized nonparametric regression and its properties.**

Keywords: nonparametric regression, collective estimation, simplified approximations, asymptotic properties.

The technique of synthesis and analysis of the generalized nonparametric regression for solution of problems of restoration of stochastic associations under conditions of the incomplete information is proposed. The idea of the considered approach consists in construction of a set of the simplified parametrical models concerning system of reference points of a learning sample. Their subsequent integration into the generalized model is carried out on the basis of methods of a nonparametric statistician. Properties of the generalized nonparametric regression are investigated. Results of its comparison with a traditional nonparametric regression are analyzed. Conditions of their competence are defined.