УДК 519.7

© 2017 г. **А.В. Лапко** [1,2], д-р техн. наук,
В.А. Лапко [1,2], д-р техн. наук
([1]Институт вычислительного моделирования СО РАН, Красноярск,
[2]Сибирский государственный аэрокосмический университет имени академика М.Ф. Решетнева, Красноярск)

НЕПАРАМЕТРИЧЕСКИЙ КОЛЛЕКТИВ ЛИНЕЙНЫХ АППРОКСИМАЦИЙ В ЗАДАЧЕ ВОССТАНОВЛЕНИЯ СТОХАСТИЧЕСКИХ ЗАВИСИМОСТЕЙ

Рассматривается методика построения непараметрического коллектива в задаче восстановления стохастических зависимостей, обеспечивающего эффективный учет априорной информации исходных статистических данных. Проводится анализ структуры непараметрического коллектива и его свойства.
Ключевые слова: непараметрическая регрессия, линейные аппроксимации, коллектив решающих функций, асимптотические свойства.

DOI: 10.22250/isu.2017.52.64-71

Введение

Обязательным условием синтеза традиционных моделей коллективного типа является наличие конечного множества решающих функций, каждая из которых имеет самостоятельное значение. Тогда коллектив моделей, – например, с позиций «средневзвешенного» преобразования либо оценивания областей их компетентности, – аккумулирует преимущества составляющих коллектив решающих функций [1]. Другим крайним случаем коллектива являются непараметрические модели, структуру которых образуют элементы обучающей выборки и соответствующие им ядерные функции. Каждая ядерная функция оказывает влияние на процесс формирования решения только в пределах конкретной ситуации из обучающей выборки [2 – 4].

Исследуемые непараметрические коллективы решающих функций формируются на основе семейства упрощенных параметрических аппроксимаций искомой зависимости, объединение которых в единую модель осуществляется с помощью непараметрической оценки условного математического ожидания относи-

тельно аргументов восстанавливаемой зависимости. Впервые данная идея была представлена в работе [5]. В нашей статье методы синтеза и анализа подобных непараметрических коллективов линейных аппроксимаций развиваются и систематизируются.

Синтез непараметрических коллективов линейных решающих функций

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

$$y = \varphi(x) \,\forall\, x \in R^k \tag{1}$$

и ее аргументов x^i. Считается, что функция (1) и плотности вероятности $p(x)$, $p(x, y)$ достаточно гладкие и имеют хотя бы две производные.

Структуру изучаемого класса моделей составляет множество упрощенных параметрических аппроксимаций исследуемой функции (1), каждая из которых строится относительно некоторой системы «опорных» ситуаций из обучающей выборки. Объединение упрощенных аппроксимаций в коллектив осуществляется с помощью непараметрической оценки оператора условного математического ожидания относительно «опорных» ситуаций.

Поставим в соответствие некоторым точкам обучающей выборки $\left(x^i, y^i\right)$ упрощенные параметрические аппроксимации $\varphi_i\left(x, \alpha^i\right)$ (опорные функции) зависимости (1), параметры которых удовлетворяют условиям

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

$$\overline{\alpha}^i = \arg\min \frac{1}{n-1} \sum_{\substack{j=1 \\ j \neq i}}^{n} (y^j - \varphi_i(x^j, \alpha^i))^2,\ i \in I_0 \subset I, \tag{2}$$

где I – множество номеров элементов исходной выборки V.

Каждая i-я упрощенная аппроксимация проходит через i-ю «опорную» точку и близка в среднеквадратическом ко всем остальным элементам обучающей выборки.

Упрощенные параметрические аппроксимации $\varphi_i(x, \alpha^i)$ могут быть линейными. В этом случае

$$\varphi_i(x, \alpha^i) = \sum_{v=1}^{k} \alpha_v^i x_v + \beta^i, \tag{3}$$

где параметры $\beta^i = y^i - \sum_{v=1}^{k} \alpha_v^i x_v^i$, а коэффициенты $\alpha_v^i,\ v = \overline{1,k}$ находятся из условия

минимума критерия $\sum\limits_{\substack{j=1\\ j\neq i}}^{n}\left[(y^j-y^i)-\sum\limits_{v=1}^{k}\alpha_v^i(x_v^j-x_v^i)\right]^2$.

Тогда задача определения параметров α может быть сведена к решению системы линейных уравнений

$$\alpha_t^i\sum_{\substack{j=1\\ j\neq i}}^{n}(x_t^j-x_t^i)^2+\sum_{\substack{v=1\\ v\neq t}}^{k}\alpha_v^i\sum_{\substack{j=1\\ j\neq i}}^{n}(x_v^j-x_v^i)(x_t^j-x_t^i)=\sum_{\substack{j=1\\ j\neq i}}^{n}(y^j-y^i)(x_t^j-x_t^i),\ t=\overline{1,k},$$

относительно $\alpha_t^i, t=\overline{1,k}$ (используя, например, правило Крамера либо метод Гаусса).

Объединение упрощенных параметрических аппроксимаций в коллектив осуществляется на основе процедуры условного усреднения

$$\overline{y}=\overline{\varphi}(x)=\sum_{i\in I_0}\varphi_i(x,\overline{\alpha}^i)\,\lambda^i(x), \qquad (4)$$

где положительная (ограниченная значением единицы) функция $\lambda^i(x)$ определяет «вес» упрощенной линейной модели $\varphi_i(x,\overline{\alpha}^i)$ при формировании оценки искомой функции (1) в ситуации x.

Примером $\lambda^i(x)$ является функция, сформированная из евклидовых расстояний между точками (x, x^i), $\lambda^i(x)=\dfrac{\left(\sum\limits_{v=1}^{k}(x_v-x_v^i)^2\right)^{-\frac{1}{2}}}{\sum\limits_{i\in I_0}\left(\sum\limits_{v=1}^{k}(x_v-x_v^i)^2\right)^{-\frac{1}{2}}}$, либо «весовая» функция

$$\lambda^i(x)=\frac{\prod\limits_{v=1}^{k}\Phi\left(\dfrac{x_v-x_v^i}{c_v}\right)}{\sum\limits_{i\in I_0}\prod\limits_{v=1}^{k}\Phi\left(\dfrac{x_v-x_v^i}{c_v}\right)}, \qquad (5)$$

составленная из «ядерных» функций $c_v^{-1}\,\Phi\left(\dfrac{x_v-x_v^i}{c_v}\right)$, удовлетворяющих условиям положительности, симметричности и нормированности [6].

С целью повышения аппроксимационных свойств непараметрических моделей коллективного типа в условиях большого уровня зашумленности и наличия выбросов в исходных экспериментальных данных возникает задача дополнительного сглаживания модели (4) восстанавливаемой зависимости (1). Предлагается учитывать статистические оценки эффективности $\overline{W}^i$ линейных аппроксимаций

$\varphi_i\left(x, \overline{\alpha}^i\right),\ i \in I_0$. В качестве показателя эффективности i-й аппроксимации может выступать среднеквадратический критерий $\overline{W}^i = \frac{1}{n-1}\sum\limits_{\substack{j=1 \\ j\neq i}}^{n}\left(y^j - \varphi_i\left(x^j, \overline{\alpha}^i\right)\right)^2,\ i \in I_0.$

Учет показателей эффективности упрощенных аппроксимаций целесообразно осуществить путем введения в коллективную модель (4) ядерной меры близости между значением $\overline{W}^i$ и ее минимально возможным значением. Полученная модификация непараметрического коллектива линейных аппроксимаций имеет вид $\overline{y} = \overline{\varphi}(x) = \frac{\sum\limits_{i\in I_0}\varphi_i\left(x, \overline{\alpha}^i\right)\prod\limits_{v=1}^{k}\Phi\left(\frac{x_v - x_v^i}{c_v}\right)\Phi\left(\frac{0-\overline{W}^i}{c_w}\right)}{\sum\limits_{i\in I_0}\prod\limits_{v=1}^{k}\Phi\left(\frac{x_v - x_v^i}{c_v}\right)\Phi\left(\frac{0-\overline{W}^i}{c_w}\right)}$, где c_w – параметр ядерной функции $\Phi\left(\frac{0-\overline{W}^i}{c_w}\right)$, который характеризует область ее определения.

Для синтеза непараметрической модели типа (4) предлагается итерационная процедура формирования системы опорных точек. Пусть $\varphi_j\left(x, \alpha^j\right),\ j \in I_0(t)$ – некоторая система линейных аппроксимаций зависимости (1), построенная относительно «опорных» точек $\left(x^j, y^j,\ j \in I_0(t)\right)$. При этом эмпирическая ошибка расхождения между экспериментальными данными и строящейся непараметрической моделью $\Psi_t\left(\varphi_j\left(x, \alpha^j\right),\ j \in I_0(t)\right),\ \overline{W}\left(\Psi_t(\cdot)\right) = \frac{1}{\left|I_{\bar{t}}\right|}\sum\limits_{i\in I_{\bar{t}}}\left(y^i - \Psi_t\left(\varphi_j\left(x^i\right),\ j \in I_0(t)\right)\right)^2$, где $I_{\bar{t}} = I \setminus I_0(t)$ – множество номеров точек, не входящих в число «опорных» $I_0(t)$, а $\left|I_{\bar{t}}\right|$ – их количество.Вклад слагаемых в формирование эмпирической ошибки неравнозначный. Если модель $\Psi_t(\cdot)$ в некоторой точке x^i имеет максимальное расхождение с экспериментальным значением y^i, то естественно было бы принять точку $\left(x^i, y^i\right)$ в качестве «опорной» при построении $(t+1)$-й упрощенной аппроксимации.

В качестве первой «опорной точки» выбирается j-й элемент из исходной выборки V с максимальным значением y^j при условии, что y^j не является ошибкой контроля. Вторая «опорная» точка соответствует ситуации $\left(x^\lambda, y^\lambda\right)$, обеспечивающей выполнение соотношения $\max\limits_{i\in I\setminus(j)}\left|y^i - \varphi_j\left(x^i, \overline{\alpha}^j\right)\right|$.

Процесс формирования линейных аппроксимаций заканчивается, если ошибка $\overline{W}\left(\Psi_t(\cdot)\right)$ не превышает заданного пользователем порога.

Анализ непараметрического коллектива

Проведем преобразование непараметрической модели (4) для одномерных случайных величин x, y с учетом условия $\beta^i = y^i - \alpha^i x^i$ и оптимального значения α^i, определяемого выражением $\overline{\alpha}^i = \dfrac{\sum\limits_{\substack{j=1 \\ j\neq i}}^{n} (y^j - y^i)(x^j - x^i)}{\sum\limits_{\substack{j=1 \\ j\neq i}}^{n} (x^j - x^i)^2}$.

Подставляя значения β^i и $\overline{\alpha}^i$ в непараметрическую модель (4) с учетом $\lambda^i(x)$ (5), получим

$$\overline{y} = \frac{\sum\limits_{i\in I_0} y^i \Phi\left(\dfrac{x - x^i}{c}\right)}{\sum\limits_{i\in I_0} \Phi\left(\dfrac{x - x^i}{c}\right)} + \overline{z}(x), \qquad (6)$$

где $\overline{z}(x) = \dfrac{\sum\limits_{i\in I_0} \overline{\alpha}^i \left(x - x^i\right) \Phi\left(\dfrac{x - x^i}{c}\right)}{\sum\limits_{i\in I_0} \Phi\left(\dfrac{x - x^i}{c}\right)}$.

Нетрудно заметить, что первое слагаемое в (6) представляет собой непараметрическую регрессию, обладающую свойствами асимптотической сходимости к условному математическому ожиданию $\hat{y} = M(y/x)$ – оптимальной модели (1) в среднеквадратическом смысле.

Второе слагаемое $\overline{z}(x)$ играет роль поправочного члена, значение которого снижается по мере роста объема исходной информации. Вид $\overline{z}(x)$ зависит от «опорных» функций. Поэтому, с учетом вышесказанного, выбор конкретной системы $\varphi_i(x, \alpha^i)$, $i \in I_0$ не оказывает принципиального влияния на свойства статистики $\overline{y} = \overline{\varphi}(x)$. Наличие поправочного члена делает коллектив (6) схожим с гибридными моделями [7], а слабая зависимость его свойств от видов опорных функций – с непараметрической регрессией [8, 9].

Из анализа модели типа (6) следует, что непараметрический коллектив линейных аппроксимаций учитывает не только информацию, содержащуюся в точках исходных данных (первое слагаемое), но и во взаимосвязи между ними (второе слагаемое).

Для последующего анализа коллектива (4) подставим линейный полином

вида (3) в выражение (4), получим $\bar{y} = \sum_{v=1}^{k} x_v \sum_{j \in I_0} \lambda^j(x)\,\bar{\alpha}_v^j + \sum_{j \in I_0} \lambda^j(x)\,\bar{\beta}^j$, где $\bar{\beta}^j$ – оценка свободного члена β^j линейной аппроксимации $\varphi_j(x, \alpha^j)$.

Очевидно, что исследуемая модель представляет собой линейную аппроксимацию с нелинейными коэффициентами, поэтому появляется возможность по их величинам оценивать вклад аргументов в формирование значений восстанавливаемой зависимости в конкретных условиях x_v, $v = \overline{1, k}$.

Свойства непараметрического коллектива

Предположим, что плотность вероятности $p(x)$ одномерной случайной величины x известна. В этом случае непараметрический коллектив типа (4) запишется в виде

$$\bar{\varphi}(x) = \frac{1}{Ncp(x)} \sum_{i \in I_0} y^i \Phi\left(\frac{x - x^i}{c}\right) + \frac{1}{Ncp(x)} \sum_{i \in I_0} \bar{\alpha}^i \left(x - x^i\right) \Phi\left(\frac{x - x^i}{c}\right), \qquad (7)$$

где N – количество элементов множества I_0.

Справедливо следующее утверждение.

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 обобщенной 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) объем исходных статистических данных $n \to \infty$; 4) последовательность $c = c(N) \to 0$ при $N \to \infty$, а $Nc \to \infty$.

Тогда:

смещение

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

квадратическое отклонение

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

$$\left. \times \left(\frac{A(x, y)}{4 p(x) D(x)} + A_1(x, y) \right) \right], \qquad (9)$$

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 то-

чек; $\|\Phi(u)\|^2 = \int \Phi^2(u)du$. Здесь и далее бесконечные пределы интегрирования опускаются.

Из выражений (8), (9) при выполнении условий теоремы следует асимптотическая несмещенность, сходимость в среднеквадратическом и состоятельность непараметрического коллектива (7).

Для доказательства данных утверждений используется технология преобразований, предложенная В.А. Епанечниковым при исследовании асимптотических свойств непараметрической оценки плотности вероятности [10] и развитая в работах [11, 12].

Установлено, что асимптотические свойства исследуемого непараметрического коллектива несущественно зависят от объема выборки, используемой при идентификации упрощенных аппроксимаций, и их вида. Эффективность предлагаемой модели в основном определяется законом распределения системы опорных точек и их количеством. На основе сравнения полученных показателей асимптотических свойств коллектива линейных аппроксимаций и традиционной непараметрической регрессии появляется возможность разработки критериев оценивания условий их компетентности.

Из результатов вычислительных экспериментов следует, что исследуемая статистика обладает преимуществом над традиционной непараметрической регрессией при относительно малых объемах исходных данных. При больших значениях n эффективность сравниваемых моделей сопоставима, что согласуется с результатами аналитических исследований.

Заключение

Непараметрический коллектив линейных аппроксимаций в задаче восстановления стохастических зависимостей занимает промежуточное положение между локальными и параметрическими моделями и использует их преимущества. Структуру изучаемого класса моделей составляют семейство упрощенных параметрических аппроксимаций исследуемой функции, каждая из которых строится относительно одной из «опорных» ситуаций из обучающей выборки. Объединение упрощенных аппроксимаций осуществляется с помощью непараметрической оценки оператора условного математического ожидания. Применение непараметрического коллектива обеспечивает учет не только информации, содержащейся в точках исходных статистических данных, но и во взаимосвязи между ними.

Установлено, что асимптотические свойства обобщенной непараметрической регрессии «слабо» зависят от вида упрощенных аппроксимаций и объема выборки в задаче их идентификации. Эффективность предлагаемых моделей в значительной степени определяется законом распределения системы опорных то-

чек и их количеством. Результаты исследований асимптотических свойств непараметрических коллективов линейных аппроксимаций являются основой оценивания условий их компетентности по сравнению с традиционной непараметрической регрессией. Предлагаемый непараметрический коллектив представляет собой линейную аппроксимацию с нелинейными коэффициентами, что открывает возможность разработки методики оценивания вклада аргументов в формирование значений восстанавливаемой функции.

ЛИТЕРАТУРА

1. Растригин Л.А. Гибридное распознавание // Автоматика и телемеханика. – 1993. – № 4. – С.3-20.
2. *Лапко А.В., Лапко В.А.* Коллектив непараметрических решающих функций в двуальтернативной задаче распознавания образов // Системы управления и информационные технологии. – 2009. – Т. 37, № 3.1. – С.156-160.
3. *Лапко А.В., Лапко В.А.* Непараметрическая оценка уравнения разделяющей поверхности в условиях больших выборок и ее свойства // Системы управления и информационные технологии. – 2010. – Т. 39, № 1.2. – С.300-304.
4. *Лапко А.В., Лапко В.А.* Анализ непараметрических алгоритмов распознавания образов в условиях пропуска данных // Автометрия. – 2008. – Т. 44, № 3. – С.65-74.
5. *Лапко В.А.* Синтез и анализ непараметрических моделей коллективного типа // Автометрия. – 2001. – № 6. – С.98-106.
6. *Parzen E.* On estimation of a probability density function and mode // Ann. Math. Statistic. – 1962. – Vol. 33, N 3. – P.1065-1076.
7. *Лапко А.В., Лапко В.А.* Гибридные системы распознавания образов в условиях неоднородных данных // Информатика и системы управления. – 2016. – №1(47). – С.73-81.
8. *Надарая Э.А.* Непараметрические оценки кривой регрессии // Тр. ВЦ АН ГССР. – 1965. – Вып.5. – С.56-68.
9. *Härdle W., Müller S., Sperlich M., Werwatz A*. Nonparametric and semiparametric models. – New York: Springer – Verlag, 2004.
10. *Епанечников В.А.* Непараметрическая оценка многомерной плотности вероятности // Теория вероятности и ее применения. – 1969. – Т. 14, №1. – С.156-161.
11. *Лапко А.В., Лапко В.А.* Регрессионная оценка плотности вероятности и ее свойства // Системы управления и информационные технологии. – 2012. – Т. 49, № 3.1. – С.152-156.
12. *Лапко А.В., Лапко В.А.* Свойства непараметрической оценки плотности вероятности многомерных случайных величин в условиях больших выборок // Информатика и системы управления. – 2012. – № 2(32). – С.121-126.

E-mail:
Лапко Александр Васильевич – lapko@icm.krasn.ru;
Лапко Василий Александрович – lapko@icm.krasn.ru.