РОССИЙСКАЯ АКАДЕМИЯ НАУК
СИБИРСКОЕ ОТДЕЛЕНИЕ
А В Т О М Е Т Р И Я

№ 6 2001

УДК 519.7

В. А. Лапко

(Красноярск)

СИНТЕЗ И АНАЛИЗ НЕПАРАМЕТРИЧЕСКИХ МОДЕЛЕЙ КОЛЛЕКТИВНОГО ТИПА*

Предлагается методика синтеза и анализа непараметрических моделей коллективного типа при решении задач восстановления стохастических зависимостей в условиях неполной информации. Идея рассматриваемого подхода состоит в построении упрощенных параметрических моделей относительно некоторого набора точек из обучающей выборки с последующей их организацией в коллектив на основе методов непараметрической статистики. Исследуются свойства полученных непараметрических аппроксимаций, анализируются результаты их сравнения с непараметрической регрессией. Решаются проблемы оптимизации непараметрических моделей и оценивания условий их компетентности.

Введение. Принципы коллективного оценивания находят широкое распространение на современном этапе развития теории обучающихся систем, когда возникла потребность обобщения разнотипных методов с целью получения интегрированных знаний.

Обязательным условием синтеза традиционных моделей коллективного типа является наличие конечного множества решающих правил, каждое из которых имеет самостоятельное значение. Тогда коллектив моделей, например, с позиций «средневзвешенного» преобразования либо оценивания областей их компетентности аккумулирует преимущества составляющих коллектив решающих правил. Другим примером 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.

В настоящее время настойчиво обсуждается и разрабатывается идея о совместном использовании в коллективе 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], полупараметрические и частично линейные модели [1]. При этом особое внимание уделяется алгоритмам восстановления стохастических зависимостей, обеспечивающих учет частичных сведений об их виде и данных экспериментальных исследований.

* Работа выполнена при поддержке Российского фонда фундаментальных исследований (проекты № 00-01-00001, № 01-01-06015).

Предлагаются непараметрические модели стохастических зависимостей коллективного типа, позволяющие в наиболее полном объеме использовать информацию обучающих выборок на основе управляемого сочетания преимуществ параметрических и локальных аппроксимаций восстанавливаемой функции.

Структуру изучаемого класса моделей составляет множество упрощенных параметрических аппроксимаций исследуемой функции, 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.

Подобные модели адекватны уровню априорной неопределенности, соответствующему традиционным непараметрическим аппроксимациям, и обобщают их.

В работе приводится методика оптимального синтеза и анализа непараметрических коллективов в задаче восстановления стохастических зависимостей, исследуются их асимптотические свойства и рассматриваются вычислительные аспекты применения моделей.

1. **Непараметрические модели коллективного типа.** Пусть дана выборка $V=(x^i,\ y^i,\ i=\overline{1,n})$ 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.

Поставим в соответствие некоторой точке $(x^i,\ y^i)$ обучающей выборки V аппроксимацию $\varphi_i(x,\overline{\alpha})$ зависимости (1), параметры α которой удовлетворяют условиям:

$$y^i=\varphi_i(x^i,\overline{\alpha}^i),\qquad \overline{\alpha}^i=\underset{\alpha^i}{\operatorname{argmin}}\frac{1}{n-1}\sum_{\substack{j=1\\ j\neq i}}^{n}(y^j-\varphi_i(x^j,\alpha))^2,\qquad i=\overline{1,N}. \tag{2}$$

Упрощенные аппроксимации $\varphi_i(x,\overline{\alpha}^i)$ проходят через опорные точки $(x^i,\ y^i,\ i=\overline{1,N})$ и близки в среднеквадратическом к остальным элементам обучающей выборки V. Для линейных упрощенных аппроксимаций

$$\varphi_i(x,\overline{\alpha})=\sum_{\mathrm{v}=1}^{k}\alpha^i_{\mathrm{v}}x_{\mathrm{v}}+\beta^i.$$

Параметр $\beta^i=y^i-\sum_{\mathrm{v}=1}^{k}\alpha^i_{\mathrm{v}}x^i_{\mathrm{v}}$, а коэффициенты α^i_{v}, $\mathrm{v}=\overline{1,k}$, находятся из условия минимума критерия

$$\sum_{\substack{j=1\\ j\neq i}}^{n}\left[(y^j-y^i)-\sum_{\mathrm{v}=1}^{k}\alpha^i_{\mathrm{v}}(x^j_{\mathrm{v}}-x^i_{\mathrm{v}})\right]^2.$$

В одномерном случае оценки значений искомых параметров

$$\bar{\alpha}^i = \frac{\sum\limits_{\substack{j=1 \\ j\neq i}}^{n}(y^j - y^i)(x^j - x^i)}{\sum\limits_{\substack{j=1 \\ j\neq i}}^{n}(x^j - x^i)^2}, \quad i=\overline{1,N}. \tag{3}$$

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

$$\bar{y} = \bar{\varphi}(x) = \sum_{i=1}^{N}\varphi_i(x,\bar{\alpha}^i)\lambda^i(x), \tag{4}$$

где положительная, ограниченная значением единица функция $\lambda^i(x)$ определяет «вес» правила $\varphi_i(x,\bar{\alpha}^i)$ при формировании решения в ситуации x. Примером функции $\lambda^i(x)$ является нормированное расстояние между точками (x,x^i) либо «весовая» функция

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

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 [3].

Непараметрическая модель коллективного типа (4) допускает представление $\bar{y} = \tilde{\varphi}(x) + \bar{z}(x)$, где первое слагаемое $\tilde{\varphi}(x)$ является непараметрической регрессией, построенной по опорным точкам, а второе $\bar{z}(x)$ играет роль поправочного члена и отражает условную взаимосвязь между точками обучающей выборки, значения которого снижаются по мере роста объема исходной информации. Наличие поправочного члена делает коллектив (4) схожим с гибридными моделями, а слабая зависимость его свойств от вида опорных функций – с непараметрической регрессией. Для одномерного случая непараметрический коллектив принимает вид

$$\bar{y}(x) = \frac{\sum\limits_{i=1}^{N} y^i\Phi\left(\frac{x - x^i}{c}\right)}{\sum\limits_{i=1}^{N}\Phi\left(\frac{x - x^i}{c}\right)} + \bar{z}(x), \qquad \bar{z}(x) = \frac{\sum\limits_{i=1}^{N}(x - x^i)\lambda^i(x)\sum\limits_{\substack{j=1 \\ j\neq i}}^{n}(y^j - y^i)(x^j - x^i)}{\sum\limits_{\substack{j=1 \\ j\neq i}}^{n}(x^j - x^i)^2}.$$

Таким образом в разрабатываемой непараметрической модели коллективного типа осуществляется двойное сглаживание: при оценивании параметров упрощенных аппроксимаций и за счет использования оператора условного математического ожидания, что обеспечивает высокую помехозащищенность принимаемых решений. При формировании решения участвуют упрощенные аппроксимации восстанавливаемой зависимости с весами, определяемыми «ядерной» мерой близости между контрольной и «опорными» ситуациями. Применение непараметрических коллективов позволяет использовать не только информацию о локальном поведении восстанавливаемой зависимости, но и вскрывать ее относительные интегральные свойства, содержащиеся в обучающей выборке.

2. **Асимптотические свойства непараметрических коллективов.** Для удобства последующего анализа предположим, что x – скаляр и закон распределения $p(x)$ известен. Тогда непараметрическая модель коллективного типа принимает вид

$$\bar{y} = \frac{1}{Ncp(x)} \sum_{i=1}^{N} \varphi_i(x, \bar{\alpha}^i) \Phi\left(\frac{x - x^i}{c}\right). \tag{5}$$

Запишем оценку непараметрической модели коллективного типа (5) с учетом выражения (3) и условия (2) прохождения через опорную точку в виде статистики

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

которая позволяет упростить методику исследования асимптотических свойств $\bar{y}$.

Теорема. Пусть

а) $\varphi(x)$, $p(x, y)$ и $p(x)$ в области определения $y = \varphi(x)$ являются ограниченными и непрерывными со всеми своими производными до второго порядка включительно;

б) ядерные функции $\Phi(u)$ являются положительными, симметричными и нормированными при $\int u^m \Phi(u) du < \infty \; \forall \; m < \infty$;

в) последовательность $c = c(n) \to 0$ при $n \to \infty$, а $Nc \to \infty$;

г) количество упрощенных аппроксимаций $N \to \infty$.

Тогда непараметрическая модель коллективного типа $\bar{y} = \bar{\varphi}(x)$ обладает свойствами асимптотической несмещенности и состоятельности.

Асимптотические выражения смещения оценки (5) и ее среднеквадратического отклонения после стандартных аналитических преобразований принимают вид

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

$$M(\bar{y}(x) - y(x))^2 \sim \frac{y^2(x) \left\| \Phi(u) \right\|^2}{Ncp(x)} + \tag{6}$$

$$+\frac{c^4}{p(x)}\left[\frac{((y(x)p(x))^{(2)})^2}{4p(x)}+\frac{A(x,y)}{D(x)}\left(\frac{A(x,y)}{4p(x)D(x)}+A_1(x,y)\right)\right], \qquad (7)$$

где $A(x,y)$, $A_1(x,y)$ – нелинейные функционалы от $\varphi(x)$, $p(x,y)$, $p(x)$ и их производных; $D(x)$ – дисперсия опорных точек; $\|\Phi(u)\|^2=\int\Phi^2(u)du$.

Из асимптотических выражений (6), (7) при $c\to 0$ и $Nc\to\infty$ следует асимптотическая несмещенность и сходимость в среднеквадратическом непараметрической модели коллективного типа $\bar{y}$. Установлено, что асимптотические свойства непараметрических моделей коллективного типа *слабо* зависят от вида упрощенных аппроксимаций и объема выборки в задаче их идентификации. Эффективность предлагаемых моделей в значительной степени определяется законом распределения системы опорных точек и их количеством. Данные выводы подтверждает выражение минимального среднеквадратического отклонения при оптимальном значении параметра размытости $c(n)$:

$$M(\bar{y}(x)-y(x))^2=\frac{5}{4}\left[\left(\frac{y^2(x)\|\Phi(u)\|^2}{Np(x)}\right)^4\left(\frac{(y(x)p(x))^{(2)}}{p(x)}\right)^2\times\right.$$

$$\left.\times\frac{A(x,y)(A(x,y)+4p(x)D(x)A_1(x,y))}{D^2(x)}\right]^{1/5}. \qquad (8)$$

3. **Оптимизация непараметрических моделей коллективного типа.** Оптимизация непараметрических моделей коллективного типа охватывает выбор оптимального закона распределения опорных точек и их количества.

Предлагаются два подхода к формированию последовательности опорных точек, основанных на их статистическом моделировании с рациональным законом распределения и итерационной процедуре выбора упрощенных аппроксимаций, минимизирующих на каждом этапе относительную эмпирическую ошибку расхождения между восстанавливаемой зависимостью и ее коллективной моделью.

Выбор рационального закона распределения опорных точек. Оптимальный закон распределения системы опорных точек находится из условия минимума среднеквадратического критерия (8) при оптимальном коэффициенте размытости.

Вследствие сложности задачи будем искать рациональный закон распределения опорных точек в виде смеси плотностей вероятностей, минимизирующих интегральные характеристики отдельных частей критерия (8). В результате можно получить последовательность вариационных задач. Вариационная задача, соответствующая первому члену выражения (8),

$$\min_{p(x)}\int\frac{y^2(x)}{p(x)}dx, \qquad \int p(x)dx=1,$$

дает решение

$$p_1^*(x)=\frac{y(x)}{\int y(x)dx}.$$

Вариационная задача для второго члена произведения в (8) имеет вид

$$\min_{p(x)} \int \left(\frac{(y(x)p(x))^{(2)}}{p(x)} \right)^2 dx, \qquad \int p(x)dx = 1,$$

решение которой соответствует

$$p_2^*(x) = \frac{\lambda}{\|\lambda\|},$$

где $\lambda = 2(y^{(1)}y^{(2)})^{(1)} - (yy^{(2)})^{(1)} - yy^{(2)}$; $\|\lambda\| = \int \lambda dx$. Из анализа третьего члена произведения в критерии (8) следует, что его значение снижается с ростом дисперсии опорных точек $D(x)$. Этому способствует использование равномерного закона распределения опорных точек $p(x)$.

Таким образом, для снижения смещения непараметрических аппроксимаций коллективного типа рациональным законом распределения является смесь плотностей вероятностей, состоящая из равномерной плотности и плотностей, зависящих от восстанавливаемой зависимости и ее производных. Опорные точки рекомендуется выбирать в области больших значений восстанавливаемой зависимости, так как при больших значениях функции ошибки имеют большую величину. Ошибки также велики в области резкого изменения значений функции, что соответствует большим значениям производной искомой функции.

Итерационная процедура формирования опорных точек. Идея предлагаемого метода основывается на процедуре последовательного формирования упрощенных аппроксимаций $(\varphi_j(x), j = \overline{1,t})$, минимизирующих на каждом t-м этапе относительную эмпирическую ошибку расхождения между восстанавливаемой зависимостью и ее коллективной моделью $\Psi_t(\varphi_j(x), j = \overline{1,t})$:

$$\overline{W}(\Psi_t(\cdot)) = \frac{1}{|I_{\bar{t}}|} \sum_{i \in I_{\bar{t}}} (y^i - \Psi_t(\varphi_j(x^i),\ j = \overline{1,t}))^2,$$

где $I_{\bar{t}} = I \setminus I_t$ – множество номеров точек, не входящих в число опорных I_t; $|I_{\bar{t}}|$ – их количество; I – множество номеров точек исходной выборки. Если модель $\Psi_t(\cdot)$ 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 $(t+1)$ упрощенной аппроксимации.

Методика синтеза модели:

а) взять в качестве первой опорной точки (x^j, y^j) наблюдение из обучающей выборки с максимальным значением восстанавливаемой функции. Принять значение параметра $t = 1$, $I_t = \{j\}$, $I_{\bar{t}} = I \setminus I_t$, $I = \{i = \overline{1,n}\}$;

б) проверить гипотезу о том, что j-е наблюдение является ошибкой (выбросом). Если j-е наблюдение ошибочное, принять $j \in I_0$ (I_0 – множество номеров ошибочных точек) и перейти к этапу е);

в) оценить параметры модели $\varphi_j(x, \alpha_j)$ из условия

$$\min_{\alpha_j} \sum_{i \in I_{\bar{t}}} (y^i - \varphi_j(x^i, \alpha_j))^2;$$

г) включить номер j-й опорной точки в множество I_{t+1}. Проверить соответствие количества опорных точек $|I_{t+1}|$ требуемому N. Если условие выполнено, то процесс заканчивается;

д) построить непараметрическую модель коллективного типа $\Psi_{t+1}(\cdot)$;

е) определить следующую опорную точку (x^j, y^j) из условия

$$\max_{x^i,\ y^i,\ i\in I_{\overline{t+1}}\setminus I_0} (y^i - \Psi_{t+1}(\varphi_{\nu}(x^i, \bar{\alpha}_{\nu}),\ \nu=\overline{1,t+1}))^2.$$

Принять $t = t+1$ и перейти к этапу б).

4. **Оценка области компетентности моделей.** В результате аналитических исследований отношения среднеквадратических критериев точности аппроксимации оценена область преимущества разработанных моделей по сравнению с непараметрической регрессией, границы которой для равномерных законов распределения аргументов зависимости и опорных точек определяются неравенством

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

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

Границы между областями компетентности моделей в новой системе координат (поворот осей) представляются эллипсом, параметры которого зависят от отношения N/n. При $N/n > 0{,}5$ наблюдается экспоненциальный рост параметров, определяющих область компетентности моделей. По данным вычислительного эксперимента при $N \geq 6k$ обеспечивается преимущество предлагаемых коллективов над непараметрической регрессией.

5. **Оценивание вклада аргументов восстанавливаемой зависимости в формирование ее значений.** Примем линейные упрощенные аппроксимации $\varphi_i(x, \alpha) = \sum_{\nu=1}^{k} \alpha^i_{\nu} + b^i$ в качестве опорных. Тогда непараметрическая модель коллективного типа (4) допускает представление в виде линейного полинома с нелинейными коэффициентами

$$\bar{y}(x) = \sum_{\nu=1}^{k} x_{\nu} \sum_{i=1}^{N} \alpha^i_{\nu} \lambda^i(x) + \sum_{i=1}^{N} b^i \lambda^i(x),$$

$$b_{\nu}(x) = \sum_{i=1}^{N} \alpha^i_{\nu} \lambda^i(x), \qquad b_0(x) = \sum_{i=1}^{N} b^i \lambda^i(x).$$

Поэтому появляется возможность по значениям его коэффициентов оценить вклад аргументов x_{ν}, $\nu = \overline{1,k}$, в формирование значений $\bar{y}(x)$.

Методика оценивания вклада аргументов:

– для обеспечения равенства интервалов изменения аргументов провести их нормировку;

– построить линейные упрощенные аппроксимации в нормированном пространстве аргументов и сформировать непараметрическую модель коллективного типа:

$$\bar{y}(x) = \sum_{\nu=1}^{k} x_\nu b_\nu(x) + b_0(x);$$

– при конкретных x рассчитать значения

$$b'_\nu(x) = \frac{b_\nu(x)}{\sum_{t=0}^{k} b_t(x)}, \quad \nu = \overline{0,k},$$

которые определяют вклад аргументов x_ν, $\nu = \overline{1,k}$, и свободного члена модели в формирование значения оцениваемой многомерной зависимости в ситуации x. При этом величина $b'_0(x)$ может быть интерпретирована как вклад в $y(x)$ неучтенных факторов при формировании исходного набора признаков.

Заключение. Рассмотренный в работе новый класс непараметрических моделей коллективного типа для решения задач восстановления стохастических зависимостей занимает промежуточное положение между локальными и параметрическими методами аппроксимации функций и использует их преимущества. Формально процесс синтеза моделей сводится к непараметрическому оцениванию функционалов от семейства регрессий, построенных относительно системы опорных точек из экспериментальных данных. Установлено, что асимптотические свойства непараметрических коллективов в основном определяются законом распределения и количеством «опорных» точек и не зависят от вида регрессий.

Проблема комплексной оптимизации непараметрических моделей коллективного типа охватывает оценивание условий их компетентности, выбор оптимальных законов распределения «опорных» точек модели, определение их количества и методики формирования. Система «опорных» точек с законом распределения, повторяющих восстанавливаемую зависимость, минимизирует главный член дисперсии непараметрической модели коллективного типа. В общем случае рациональный закон распределения зависит не только от значений восстанавливаемой зависимости, но и от ее производных. На основе анализа области компетентности непараметрических моделей коллективного типа предложены численные критерии выбора количества «опорных» точек по объему исходной выборки и максимальных значений восстанавливаемой зависимости и ее производной. Разработаны методики управляемого синтеза структуры непараметрической модели коллективного типа, основанные на моделировании системы «опорных» точек с рациональным законом распределения и итерационной процедуре последовательного формирования упрощенных аппроксимаций, минимизирующих на каждом этапе относительную эмпирическую ошибку моделирования. Полученные результаты обобщают традиционные непараметрические методы, основанные на оценках плотности вероятности типа Розенблатта – Парзена, и открывают новое научное направление моделирования неопределенных систем.

СПИСОК ЛИТЕРАТУРЫ

1. **Хардле В.** Прикладная непараметрическая регрессия. М.: Мир, 1993.
2. **Катковник В. Я.** Линейные и нелинейные методы непараметрического регрессионного анализа // Автоматика. 1979. № 5. С. 165.
3. **Лапко А. В., Ченцов С. В., Крохов С. И., Фельдман Л. А.** Обучающиеся системы обработки информации и принятия решений. Новосибирск: Наука, 1996.

Институт вычислительного моделирования СО РАН,
E-mail: lapko@ksc.krasn.ru

Поступила в редакцию
1 декабря 2000 г.