МАШИННЫЕ МЕТОДЫ ОБНАРУЖЕНИЯ ЗАКОНОМЕРНОСТЕЙ
(Вычислительные системы)
1981 год Выпуск 88

УДК 681.3.06:621.391

КОМБИНИРОВАННЫЕ АЛГОРИТМЫ ОБНАРУЖЕНИЯ ЗАКОНОМЕРНОСТЕЙ

А.В.Лапко

Существенным вкладом в совершенствование и развитие математического обеспечения исследования разнообразных объектов является использование методологии автоматического обнаружения закономерностей. Интерпретация процесса решения задач анализа экспериментальных данных, идентификации, управления как процесса последовательного усиления эмпирических гипотез позволяет синтезировать эффективные универсальные алгоритмы принятия решений [1,2].

Данная работа посвящена разработке алгоритмов построения математических моделей, позволяющих в наиболее полной мере использовать априорную информацию об исследуемом объекте, либо вскрывать свойства восстанавливаемых стохастических зависимостей в результате анализа экспериментальных данных. Последнее обусловлено тем, что известные стохастические аппроксимации параметрического и непараметрического вида в основном ориентированы на определенный тип данных. Так, если в параметрическом подходе к восстановлению зависимости $y = \varphi(x)$ за основу принимаются сведения $F(x,\alpha)$ (или гипотеза) о структуре преобразования $\varphi(x)$ с точностью до набора параметров α, то для непараметрических методов достаточно знания лишь статистической выборки наблюдений (x^i, y^i), $i = \overline{1,n}$, и некоторых качественных характеристик оператора $\varphi(x)$ (однозначность). В первом случае из-за сжатия выборки (x^i, y^i), $i = \overline{1,n}$, в вектор достаточных статистик $\hat{\alpha}$ (оценки параметров α) теряется полезная информация о локальном поведении зависимости $y = \varphi(x)$ и, как следствие этого, появляется систематическая ошибка. Во втором - не учитываются априорные сведения о виде преобразования $\varphi(x)$.

С другой стороны, при решении конкретных задач часто отсутствует достоверная информация о непрерывности и однозначности зависимости $y = \varphi(x)$, поэтому прямое применение параметрических и непараметрических моделей неизбежно приводит к снижению их эффективности.

Для преодолёния отмеченных недостатков предлагается использовать комбинированные алгоритмы, представляющие собой комплекс различных по назначению и методам синтеза решающих правил, состав и порядок функционирования которых в процессе построения математической модели определяется уровнем априорных данных об исследуемом объекте.

Примеры комбинированных алгоритмов.

I. Для восстановления зависимости $y = \varphi(x)\ \forall\ x \in \Omega\ (\ x\ ) \subset R^k$ по конечному числу статистически независимых наблюдений (x^i, y^i), $i = \overline{1,n}$, и априорным сведениям $\varphi(x) \in \{F(x,\alpha),\ \alpha \in R^m\}$ предлагается подход, основанный на совместном применении параметрических и непараметрических моделей. При этом сначала строится параметрическая модель $F(x,\hat{\alpha})$, а с помощью непараметрических методов оценивается функция невязки $f(x) = \varphi(x) - F(x,\hat{\alpha})$.

Тогда комбинированный алгоритм апроксимации имеет вид

$$\hat{y} = \hat{\varphi}(x) = F(x,\hat{\alpha}) + \tilde{f}(x), \qquad (1)$$

где оценивание вектора $\hat{\alpha}$ и преобразования $f(x)$ как непараметрической оценки $\tilde{f}(x)$ условного математического ожидания [3] осуществляется соответственно по выборкам (x^i, y^i), $(x^i, f(x^i))$, $i = \overline{1,n}$.

Преимущество данного подхода определяется возможностью одновременного учета как информации о локальном поведении восстанавливаемой зависимости, так и частичных сведений о ее структуре, снижением требований к точности оценивания параметров α по сравнению с параметрическими алгоритмами. Кроме того, в задачах управления использование комбинированной модели (I) позволяет облегчить поиск условных экстремумов функции $y = \varphi(x)$. Здесь алгоритм оптимизации очевиден: сначала по $F(x, \hat{\alpha})$ на основе аналитических методов находится начальная экстремальная точка x_0^*; затем в окрестности этой точки с использованием известных алгоритмов оптимизации по модели (I) ищется уточнение x_0^*.

Доказывается асимптотическая несмещенность и состоятельность оценки (I), проводится ее сравнение с параметрическими и непараметрическими алгоритмами аналитически и путем моделирования на

ЭВМ. В частности, если вектор x имеет равномерное распределение $P(x)$, то квадратичные критерии аппроксимации функции $\varphi(x)$ оценкой (I) и непараметрической регрессией $\tilde{y} = \tilde{\varphi}(x)$ (см. [3]) связаны соотношением

$$M\{(y - \hat{y})^2\} = M\{(y - \tilde{y})^2\} -$$

$$- \frac{5}{2^{9/5}} \left[F''(x,\alpha) \left(y'' - \frac{F''(x,\alpha)}{2} \right) \left(\frac{l(y^2 - f^2(x))}{np(x)} \right)^4 \right]^{1/5}, \quad l > 0.$$

2. Пусть объект описывается разрывно-неоднозначной зависимостью

$$y = \varphi(x) = \varphi_j(x) \forall (\omega = (x,y)) \in \Omega_j(\omega), \; \Omega_j(\omega) \bigcap_{j \neq s} \Omega_s(\omega) = \emptyset, \tag{2}$$

$$j,s = \overline{1,m},$$

причем вид однозначных преобразований $\varphi_j(x)$, $j = \overline{1,m}$, и m-модальной плотности вероятности $P(\omega) \forall \omega \in \Omega(\omega)$, принимающей одно из экстремальных значений в каждой области $\Omega_s(\omega)$, $s = \overline{1,m}$, не задан. Имеется выборка наблюдений $\omega^i \in \Omega(\omega)$, $i \in I = [i = \overline{1,n}]$.

Суть комбинированного алгоритма восстановления зависимости (2) на основании наблюдений $\omega^i \in \Omega(\omega)$, $i \in I$, состоит в выделении областей $\Omega_j = \Omega_j(\omega)$, $j = \overline{1,m}$, соответствующих модам (областей однозначности $y = \varphi(x)$) $p(\omega)$ с помощью алгоритмов автоматической классификации и в последующем оценивании в этих областях функций $y_j = \varphi_j(x)$, $j = \overline{1,m}$.

Для реализации I-го этапа идентификации предлагается алгоритм последовательного восстановления непараметрической оценки уравнения разделяющей поверхности $\tilde{f}_{s\bar{s}}(\omega) = \tilde{P}_s \tilde{P}_s(\omega) - \tilde{P}_{\bar{s}} \tilde{P}_{\bar{s}}(\omega)$ между областями Ω_s, $\Omega_{\bar{s}} = \Omega \setminus \Omega_s$ (см.[4]): $F^t_{s\bar{s}}(\omega^i)$: $\omega^i \in \Omega_s$, если

$$\tilde{f}^{t-1}_{s\bar{s}}(\omega^i) = \sum_{j \in I^s_{t-1}} n^{-1} \prod_{\nu=1}^{k+1} \bar{c}_\nu^{-1} \Phi\left(\frac{\omega^i_\nu - \omega^j_\nu}{\bar{c}_\nu} \right) > 0, \tag{3}$$

$$i \in \bar{I}^{s-1}_{t-1} = I \setminus (I^s_{t-1} \cup (I^{s-1} = \bigcup_{\tau=1}^{s-1} I^\tau_{P_\tau})),$$

$$I^s_{t-1} = [i: \tilde{f}^{t-2}_{s\bar{s}}(\omega^i) > 0, \; i \in I \setminus I^{s-1}], \quad t = 1, P_s,$$

при $\rho(\omega_\mu, v_\mu) = \min\limits_{\omega, v} \max\limits_{\nu=\overline{1,k+1}} |\omega_\nu - v_\nu| > \bar{c}_\mu$, $\omega \in \Omega_s$, $v \in \Omega_{\bar{s}}$, где $\bar{c}_\nu$, $\nu = \overline{1,k+1}$, – субоптимальные значения коэффициентов размытости $\tilde{f}_{s\bar{s}}(\omega)$, $s = \overline{1,m}$, определяемые из условия $\min\limits_{c_\nu} [\int \tilde{P}^2(\omega)d\omega - 2/n \sum\limits_{i \in I} \tilde{P}(\omega^i)]$.

Процесс выделения $\omega^i \in \Omega_s$ заканчивается при выполнении соотношения $\tilde{f}^{P_s}_{s\bar{s}}(\omega^i) = 0 \quad \forall i \in \bar{I}^{s-1}_{P_s}$. Для формирования выборки $\omega^i \in I^{s+1}_{P_{s+1}}$ следующего класса Ω_{s+1} среди оставшихся наблюдений ω^i, $i \in I \backslash I^s$, выбирается произвольная точка $\omega^\eta (\tilde{P}(\omega^\eta) \neq 0)$, т.е. $I^{s+1}_0 = [\eta]$, и описанный выше процесс самообучения повторяется.

На следующем этапе идентификации по полученным выборкам (x^i, y^i), $i \in I^s_{P_s}$, восстанавливаются $y_s = \varphi_s(x)$, $s = \overline{1, m}$, в виде непараметрической регрессии $\tilde{y}_s$ (см. [3]), либо $\hat{y}_s(I)$ (см. [5]).

Изложенный подход используется при построении адаптивных моделей многосвязных объектов, описывающихся системой нелинейных уравнений, часть из которых типа $\tilde{y} = \tilde{\varphi}(x)$. В этом случае алгоритмы классификации применяются не только при выделении областей однозначности, но и областей, соответствующих корням системы.

В ы ч и с л и т е л ь н ы е   а с п е к т ы. С целью повышения качества функционирования предложенных алгоритмов вводится в рассмотрение интегральная непараметрическая оценка [6] плотности вероятности $P(\omega)$, являющаяся обобщением оценки Парзена

$$\tilde{P}(\omega) = (nc)^{-1} \sum_{i=1}^{n} \int\limits_{\Omega(u)} \Phi\left(\frac{\omega - \beta u - \omega^i}{c}\right) h(u) du, \qquad (4)$$

где $\Phi(\cdot)$, $h(\cdot)$ – колоколообразные функции, удовлетворяющие условиям регулярности (см. [6]), а $c \to 0$, $\beta \to 0$, $nc \to \infty$ при $n \to \infty$. Исследуются асимптотические свойства оценки (4), вопросы ее оптимизации по c, β и $\Phi(\cdot)$.

Комбинированные алгоритмы широко использовались в задачах построения моделей рудных месторождений и при обработке аэрокосмической информации.

Л и т е р а т у р а

I. ЗАГОРУЙКО Н.Г. Искусственный интеллект и эмпирическое предсказание (спецкурс для студентов НГУ). – Новосибирск, 1975, – I20 с.

2. ЗАГОРУЙКО Н.Г. Некоторые проблемы автоматического обнаружения закономерностей. -В кн.: Машинные методы обнаружения закономерностей. Новосибирск, 1976, с.6-15.

3. НАДАРАЯ Э.А. Об оценке регрессии. - Теория вероятностей и ее применения, 1964, т.9, №1, с.157-159.

4. ЛАПКО А.В., ПЕСТУНОВ И.А. Непараметрические алгоритмы автоматической классификации образов. - Красноярск, Б.и., 1979 (Препринт/ВЦ СО АН СССР).

5. ЛАПКО А.В., ИЛЬИН Е.В., ПЕСТУНОВ И.А. Непараметрические алгоритмы автоматической классификации образов и их применение в задаче идентификации. -В кн.: Тезисы докладов УШ Всесоюзного совещания по проблемам управления. Т.2, Таллин, 1980, с. 373-375.

6. ИЛЬИН Е.В., ЛАПКО А.В. К задаче восстановления плотности вероятности в условиях непараметрической неопределенности. -Красноярск, Б.и., 1980. (Препринт/ВЦ СО АН СССР).

Поступила в ред.-изд.отд.
16 февраля 1981 года