В рассмотренной детерминистской модели вопрос о том, какая популяция выживет, зависит, прежде всего, от того, в каком направлении начинает развиваться процесс. Если принять во внимание изменения, которые могут возникнуть при стохастической формулировке этой задачи, то следует ожидать, что случайные изменения в начале процесса окажут существенное влияние на конечный результат. Интерес представляет задача о достижимости равновесного состояния для процесса, описывающего стохастическую модель роста и гибели двух конкурирующих популяций, при изменении значений характеристик.

Литература: 1. *Бейли Н.,* Математика в биологии и медицине. М.: Мир, 1970. 326 с. **2.** *Баруча-Рид А. Т.* Элементы теории марковских процессов и их приложения. М.: Наука, 1969. 512 с. **3.** *Карлин С.* Основы теории случайных процессов. М.: Мир, 1971. 536 с.

Поступила в редколлегию 26.03.2004

Рецензент: д-р техн. наук, проф. Руденко О.Г.

Яловега Ирина Георгиевна, аспирантка кафедры ПМ ХНУРЭ. Адрес: Украина, 61000, Харьков, ул. Новгородская, 20, кв. 14, тел. 30-44-41.

УДК 519.6

АНАЛИЗ АЛГОРИТМОВ ВОССТАНОВЛЕНИЯ ЗАВИСИМОСТИ

ГРИЦЮК В.И.

Рассматривается задача восстановления функции по её измерениям, содержащим случайную ошибку по независимой выборке ограниченного объема. Приближение ищется в виде отрезка ряда по заданной системе функций. Приводятся критерии подбора порядка модели, использующие ту или иную априорную информацию, и результаты экспериментального исследования эффективности критериев.

1. Введение

Решение вопроса выбора порядка модели, занимающего центральное место при выборе модельной структуры, является наиболее существенным условием, гарантирующим успешное решение прикладных задач. Приближение для восстанавливаемой функции находится в виде отрезка ряда по заданной системе функций. Мерой точности может служить среднеквадратическая ошибка прогноза функции на всем отрезке или ошибка прогноза в тех же точках. *Актуальным* является выбор алгоритма оценивания по выборке ограниченного объема. Даже в частном случае объекта конечной размерности целью решения является не отыскание истинного порядка N, а построение модели, дающей наилучший прогноз (её порядок обычно меньше N). Использование априорной информации о восстанавливаемой функции приводит к выбору подходящей модели.

Таким образом, *целью* настоящей работы является конкретизация результатов методологии восстановления зависимости, основанной на среднеквадратичной ошибке прогноза, при построении полиномиальной регрессии по выборкам конечного объема и построение на их основе реализуемых алгоритмов при различных априорных предположениях.

2. Минимизация достигнутого значения среднего риска

Для обоснованного применения на конечных выборках данных без обременительных предположений о статистических свойствах переменных известен критерий, который получен Вапником, как равномерно сходящаяся оценка сверху теоретической функции среднего риска с помощью функции эмпирического риска. Для задачи выбора структуры регрессии этот критерий может быть записан в виде

$$J(\varphi) \prec \frac{J_э(\varphi)}{1-\sqrt{\frac{N(\ln n / N+1)-\ln(1-\tau)}{n}}}, \qquad (1)$$

$J_э(\varphi)$ – функционал эмпирического риска:

$$J_э(\varphi)=\frac{1}{n}\sum_{i=1}^{n}(y_i-\varphi(x_i))^2, \qquad (2)$$

$\varphi = G(T)$ – результат оценивания с помощью алгоритма G по выборке T; N – размерность пространства, на которое отображается выборочное пространство с помощью операторов из класса π алгоритмов оценивания функции f по выборкам объёма n; τ – вероятность, с которой справедливо неравенство. Метод поиска минимума гарантированного наименьшего значения функционала среднего риска с использованием неравенства (1) описан в [1] и называется методом структурной минимизации риска. Другим способом оценки по выборке T значения функционала среднего риска является процедура, результатом применения которой есть выражение

$$W=\frac{1}{n}\sum_{i=1}^{n}(y_i-f^i(x_i))^2, \qquad (3)$$

называемое величиной скользящего контроля [1]. Известно, что величина (3) является несмещенной оценкой математического ожидания среднего риска при фиксированном алгоритме оценивания функции по независимой выборке объема n-1. Для линейных алгоритмов вместо (3) предложено выражение

$$C_v=\frac{\sum_{i=1}^{n}(y_i-a_i y)^2}{(n-SpA)^2}, \qquad (4)$$

где A – матрица размера $n\times n$, осуществляющая линейное оценивание вида $\hat{y}=Ay$, a_i – i-я строка матрицы A. Выражение (4) называется критерием взаимной значимости. Хотя величина C_v несмещена относительно математического ожидания сред-

него риска, необходимо отметить разброс критерия в отношении его среднего значения. При объёме выборки, соизмеримом с размерностью пространства N (при оценивании функции f в N-мерном функциональном пространстве), дисперсия величины C_v большая и применение последней в качестве критерия выбора алгоритма может приводить к плохим результатам.

3. Выбор порядка модели

Примем за меру точности приближения среднеквадратическую ошибку прогноза функции на всём отрезке [a,b]:

$$L_N = \frac{1}{b-a}\int_a^b (f(x)-\hat{f}_N(x))^2\,dx \qquad (5)$$

или ошибку прогноза в тех же точках x_i:

$$D_N = M\frac{1}{n}\sum_{i=1}^{n}(f(x_i)-f_N(x_i))^2 . \qquad (6)$$

Рассмотрим следующую схему наблюдений.

Имеются измерения y_i функции f(x) в точках $x_i \in [a,b], i = 1,\cdots,n$, содержащие случайную помеху

$$\xi_i,\ y_i = f(x_i)+\xi_i . \qquad (7)$$

Ищется приближение для f(x) в виде конечного ряда

$$\hat{f}_N(x) = \sum_{k=1}^{N}\hat{c}_k\varphi_k(x) . \qquad (8)$$

Помехи ξ_i не коррелированы, центрированы и имеют ограниченную дисперсию.

$$M\xi_i = 0,\ M\xi_i\xi_j = \sigma^2\delta_{ij}, i,j = 1,\dots,n .$$

Точки $x_i, i = 1,\dots,n$ фиксированы, в качестве $\hat{c}^N = (\hat{c}_1,\dots,\hat{c}_N)$ используется оценка наименьших квадратов, φ_k – заданная система функций на отрезке $[a,b] \subset R$.

Введем величину остаточной суммы квадратов

$$s_N = \frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{f}_N(x_i))^2 . \qquad (9)$$

Тогда $Ms_N = D_N + \sigma^2 - \frac{2N\sigma^2}{n}$,

$$M(s_N + 2N\sigma^2 n^{-1}) = D_N + \sigma^2,$$

поэтому N, которое минимизирует

$$C_N = s_N + 2N\sigma^2 n^{-1},$$

будет (в среднем) давать и минимальное значение D_N.

Итак, если в задаче прогноз проводится для точек x_i ("экзамен" проводится в тех же точках, что и "обучение"), то порядок модели можно выбирать из условия

$$N = \arg\min_{1\le N\le\overline{N}} C_N, C_N = s_N + \frac{2N\sigma^2}{n} . \qquad (10)$$

Этот критерий совпадает с критерием Меллоуса - Акайке. Хотя критерий (10) не требует никаких априорных сведений о задаче, но лишь для среднего C_N минимум достигается при том же N, что и для среднего риска D_N.

Кроме того, минимизировалась величина D_N, а не L_N, а в большинстве случаев интересует средний риск L_N ("экзамен" проводится на всем отрезке [a,b]) . Если доступна некоторая информация о поведении функции f(x), например, известно, что f(x) обладает той или иной гладкостью, функции $\varphi_k(x)$ -тригонометрические или алгебраические полиномы, то известна скорость убывания коэффициентов Фурье. Можно показать близость D_N и L_N при некоторых условиях к величине $A_N = N\sigma^2 n^{-1} + \mu_N$, поэтому здесь приближенно наилучший способ выбора порядка модели таков:

$$N_\cdot = \arg\min_{1\le N\le\overline{N}} A_N,\ A_N = \frac{N\sigma^2}{n} + \mu_N, \qquad (11)$$

где

$$\mu_N = \sum_{k=N+1}^{n}(c_k^*)^2 , \qquad (12)$$

$$c_k^* = (b-a)^{-1}\int_a^b f(x)\varphi(x)dx,\quad \sum_{k=1}^{\infty}(c_k^*)^2 \prec \infty, \qquad (13)$$

для случая ортогональности функций $\varphi_k(x)$ на отрезке [a,b]. Если же (12) неизвестно, то предполагая ортогональность функций $\varphi_k(x)$ на системе точек x_i, при которой выполняется

$$\begin{aligned} &\frac{1}{n}\sum_{i=1}^{n}\varphi_k(x_i)\varphi_l(x_i) = \delta_{kl}, k,l = 1,\dots,N;\\ &\varphi^N(x) = (\varphi_1(x),\dots,\varphi_N(x))^T = R^N; \end{aligned} \qquad (14)$$

$$\hat{c}^N = \frac{1}{n}\sum_{i=1}^{n} y_i\varphi^N(x_i) , \qquad (15)$$

используем $\hat{c}^N$. Если $\left|c_k^*\right| \le ck^{-\alpha}, \alpha \succ 5/2, \rho_N \ne 0$ для всех N, то для приближения функции алгебраическими полиномами [2] и равномерной сетки x_i пользоваться критериями (10) и (11) можно лишь в том случае, если восстанавливаемая функция обладает достаточно высокой гладкостью. Граница $\overline{N}$ в (10) и (11) должна удовлетворять условию

$$\overline{N} = o(n^{2/5}) . \qquad (16)$$

Например,

$$\overline{N} = cn^{1/3} . \qquad (17)$$

Исследуем результаты экспериментальной проверки эффективности различных критериев при построении полиномиальной регрессии по выборкам конечного объема.

4. Экспериментальное исследование различных критериев выбора оптимального адгоритма

Для экспериментального исследования эффективности описанных подходов была выбрана задача оценивания одномерной функции регрессии. При этом случайная выборка $T = \{x_1, y_1, \ldots, x_n, y_n\}$ генерировалась согласно модели (7) с параметрами

$$f(x) = \sin^2 \pi x, x \in [-1,1], \quad \xi \sim N(0,01), \quad n = 21, \quad x_i = 0,1(i-11). \tag{18}$$

Если определить численным интегрированием коэффициенты c_k^* с $\varphi_k(x)$ полиномами Лежандра, то можно вычислить D_N, L_N, A_N для этой задачи и оптимальные по этим критериям N. Так как эти данные неизвестны, то воспользуемся критерием (10). Для вычисления s_N введем полиномы Чебышева, ортогональные на сетке x_i, можно вместе с методом [3] ортогонального разложения. Оценки $\hat{c}^N$ имеют вид (15). Величина $s_N = n^{-1}\sum_{i=1}^{n}(y_i - (\hat{c}^N, \overline{\varphi}^N))^2$ будет той же, что и для полиномов Лежандра.

Результаты вычислений приведены в табл.1 по критерию (11) с оценками $\overline{\mu}$, критерию (1) с

$$C_N = s_N / (1 - \sqrt{n^{-1}(N \ln(n/N+1) + 3},$$

где надежность равномерной оценки τ принята равной 0,95, по критерию (4) с $C_N = \frac{s_N}{(1-\frac{N}{n})^2}$ и критерию(10).Приведены результаты вычислений по другим известным критериям (FPE, Гаусса соответственно):

Таблица 1

N п/п	Критерии					
	(11)	(1)	(4)	(10)	(19)	(20)
1	0	0	0	0	0	0
2	0	0	0	0	0	0
3	0	0	0	0	0	0
4	0	0	0	0	0	0
5	0	0	0	0	0	0
6	0	0	0	0	0	0
7	0	30	5	6	2	0
8	0	0	0	0	0	0
9	16	0	6	11	4	5
10	3	0	4	4	4	1
11	2	0	2	3	1	1
12	0	0	1	1	0	0
13	2	0	2	2	0	0
14	1	0	1	1	1	1
15	3	0	1	0	0	2
16	1	0	0	0	1	0
17	0	0	0	0	0	1
18	2	0	0	2	1	1
19	0	0	1	0	2	3
20	0	0	4	0	14	15
$\overline{N} = 20$ $\bar{D}_N$	0,0051	0,0061	0,0057	0,0054	0,0073	0,0077
$\overline{N} = 20$ $\hat{L}_N$	0,099	0,0059	24,685	0,1008	88,215	95,167
$\overline{N} = 10$ $\hat{D}_N$	0,0042	0,0061	0,0048	0,0043	0,0045	0,0057
$\overline{N} = 10$ $\bar{L}_N$	0,004	0,0059	0,0045	0,0042	0,0042	0,0055

Таблица 2

N п/п	Критерии					
	(11)	(1)	(4)	(10)	(19)	(20)
1	0	0	0	0	0	0
2	0	12	0	0	0	0
3	0	0	0	0	0	0
4	0	18	6	0	0	0
5	0	0	0	0	0	0
6	0	0	7	0	2	4
7	0	0	1	0	0	0
8	0	0	8	0	2	3
9	0	0	0	0	0	0
10	0	0	6	1	9	4
11	0	0	0	0	1	0
12	0	0	1	2	3	4
13	0	0	0	1	2	0
14	0	0	0	1	2	4
15	0	0	0	2	3	1
16	1	0	0	1	0	2
17	1	0	0	4	0	2
18	4	0	0	3	1	0
19	18	0	0	5	3	4
20	6	0	1	10	2	2
$\overline{N} = 20$ $\bar{D}_N$	0,0124	0,047	0,028	0,012	0,017	0,019
$\overline{N} = 20$ $\hat{L}_N$	89,24	0,0488	9,442	101,97	26,429	26,597
$\overline{N} = 10$ $\hat{D}_N$	0,0181	0,047	0,0175	0,0176	0,021	0,020
$\overline{N} = 10$ $\bar{L}_N$	0,022	0,0488	0,029	0,023	0,023	0,021

$$C_N = s_N (n+N)/(n-N), \quad (19)$$

$$C_N = s_N /(n-N). \quad (20)$$

Указано, сколько раз из 30 реализаций выбиралась данная степень N, а также получаемые при этом ошибки прогноза $\hat{L}_N$ и $\hat{D}_N$, т.е. величины

$$\frac{1}{2}\int_{-1}^{1}(f(x)-\hat{f}_N(x))^2 dx, \; n^{-1}\sum_{i=1}^{n}(f(x_i)-\hat{f}_N(x_i))^2,$$

усредненные по реализациям.

Видно, что применение алгебраических полиномов отразилось на поведении ряда критериев (особенно критерия Гаусса), но серьезных ошибок при их использовании не возникает.

Примем $f(x)=\begin{cases}0, x \prec 0,\\ 1, x \geq 0\end{cases}$.

Результаты вычислений по приведенным критериям даны в табл.2 для $\overline{N}$ =n-1. Видно, что применение всех критериев, кроме критерия Вапника, приводит к плохим результатам в случае интегральной ошибки прогноза. Это объясняется тем, что f(x) разрывна, коэффициенты c_k убывают медленно, значения n и σ^2 малы. Однако, если назначать верхнюю границу $\overline{N}$ меньшей n, а именно взять $\overline{N}=[4n^{1/3}]$ (в данном случае $\overline{N}$ =10), то значения N>10, приводящие к большим ошибкам , не будут выбираться.

5. Заключение

Таким образом, при восстановлении зависимости по выборкам конечного объема критерий структурной минимизации, а также критерий (11), если восстанавливаемая функция обладает достаточно высокой гладкостью ,дают примерно равные по точности результаты, близкие к оптимальным.

Назначение верхней границы $\overline{N}$ меньшей n в соответствии с соотношением (16) для приближения функции алгебраическими полиномами приводит к получению приемлемых значений ошибки прогноза при восстановлении различных зависимостей.

Литература: 1. *Михальский А. Н.* Выбор алгоритма оценивания по выборкам ограниченного объема// АиТ,1987. №7. С.91-102 **2.** *Суетин П.К.* Классические ортогональные многочлены. М.: Наука, 1976. **3.** *Грицюк В.И.* Улучшенные алгоритмы для оценки методом наименьших квадратов // Радиоэлектроника и информатика.1998. №2(3). С.64-66.

Поступила в редколлегию 22.04.2004

Рецензент: д-р техн. наук, проф. Лысенко Э.В.

Грицюк Вера Ильинична, канд. техн. наук, доцент кафедры системотехники ХНУРЭ. Научные интересы: стохастические системы управления. Хобби: литература, музыка. Адрес: Украина, 61166, Харьков, пр.Ленина, 14.