РОССИЙСКАЯ АКАДЕМИЯ НАУК
СИБИРСКОЕ ОТДЕЛЕНИЕ
АВТОМЕТРИЯ
2003, том 39, № 1

УДК 519.7

А. В. Лапко, В. А. Лапко
(Красноярск)

НЕПАРАМЕТРИЧЕСКИЕ МЕТОДИКИ АНАЛИЗА МНОЖЕСТВ СЛУЧАЙНЫХ ВЕЛИЧИН

Предлагается методика анализа множеств случайных величин в задачах восстановления стохастических зависимостей и распознавания образов, основанная на оценивании вероятностных законов распределения элементов множеств и их преобразовании с помощью непараметрических процедур. Исследованы асимптотические свойства моделей. Полученные результаты имеют актуальное значение при обработке больших массивов статистических данных.

Введение. Существующий парадокс формирования исходных данных при обнаружении статистических закономерностей состоит в сопоставлении значений случайных величин, вероятность которого невелика. Более обоснованными являются постановки задач, когда обучающие выборки составляются из пар множеств случайных величин.

Пусть состояние исследуемого объекта характеризуется множествами $X \subset R^k$ и $Y \subset R^1$ независимых наблюдений случайных величин x и y, взаимосвязь между которыми определяется неизвестным преобразованием

$$R\colon X \to Y.$$

Априорную информацию составляют N пар множеств $(X'_i, Y'_i, i=\overline{1,N})$, где множеству X'_i соответствует вполне определенное множество Y'_i. Назовем X входным, а Y выходным множеством. Подобные условия встречаются при исследовании объектов, параметры которых многократно измеряются в течение короткого интервала времени; оценивании качества партии изделий по случайно выбранному их ограниченному набору; обработке больших массивов статистических данных, которые с помощью декомпозиции преобразуются к обучающей выборке изучаемой задачи.

Предлагаются методы анализа множеств случайных величин в задачах обнаружения статистических закономерностей. Идея предлагаемого подхода заключается в замене операций над множествами менее трудоемкими и хорошо разработанными операциями над функциями либо над их параметрами с использованием методов локальной аппроксимации.

Регрессионная оценка плотности вероятности. Вычислительная эффективность непараметрических алгоритмов обработки информации во

многом определяется объемом статистических данных и снижается по мере его увеличения, что затрудняет построение систем принятия решений в условиях больших выборок.

Для обхода возникающих проблем рассмотрим методику сжатия априорной информации на основе ее декомпозиции и последующего анализа получаемых множеств случайных величин.

Пусть x^i, $i=1,n,$ – выборка из n независимых наблюдений случайной величины $x \in R^1$ с неизвестной плотностью вероятности $p(x)$, которая ограничена и непрерывна со всеми своими производными до второго порядка включительно. Причем

$$\int (p^{(2)}(x))^2 dx = \left\| p^{(2)}(x) \right\|^2 < \infty,$$

где $p^{(2)}(x)$ – вторая производная $p(x)$. Эти условия, налагаемые на $p(x)$, обозначим через G_2. Для упрощения записи бесконечные пределы интегрирования здесь и в дальнейшем опускаются.

Разобьем область определения $p(x)$ на N непересекающихся интервалов длиной 2β и сформируем множества случайных величин X^j, $j=1,N$. В качестве характеристики X^j примем частоту P^j попадания случайной величины x в j-й интервал и его центр z^j. На основе полученной информации составим статистическую выборку $(z^j,\ p^j = P^j/(2\beta),\ j=1,N)$, где центры z введенных интервалов имеют равномерный закон распределения $p(z)=(2\beta N)^{-1}$, а объем полученной выборки N может быть значительно меньше исходной n.

В качестве приближения по эмпирическим данным искомой плотности $p(x)$ примем непараметрическую оценку условного математического ожидания

$$\bar{p}(x) = c^{-1} \sum_{i=1}^{N} P^i \Phi\left(\frac{x - x^i}{c}\right), \qquad (1)$$

где $\Phi(u) \in H$ – ядерная функция, удовлетворяющая условиям регулярности H [1]:

$$\Phi(u) = \Phi(-u), \quad 0 \le \Phi(u) < \infty,$$

$$\int \Phi(u)du = 1, \quad \int u^2 \Phi(u)du = 1, \qquad (2)$$

$$\int u^m \Phi(u)du < \infty, \quad 0 \le m < \infty;$$

$c = c(n)$ – убывающая с ростом n последовательность положительных чисел (коэффициентов размытости).

Статистику (1) нетрудно получить, подставляя в выражение условного математического ожидания непараметрическую оценку совместной плотности вероятности случайных величин $(P,\ z)$ и известную плотность $p(z)$. Регрессионная оценка плотности $p(x)$ является нормированной функцией, т. е. удовлетворяет основному свойству плотности вероятности.

Асимптотические свойства $\bar{p}(x)\ \forall\, x \in R^1$ определяет

Теорема 1. Пусть $p(x)$ удовлетворяет условиям G_2, $\Phi(u) \in H$, кроме того, последовательности $c>0$, $\beta>0$ таковы, что при n, $N \to \infty$, значения $c, \beta \to 0$, $\beta/c \to 0$, $nc \to \infty$.

Тогда смещение

$$W_1(\bar{p}(x)) = M\{\bar{p}(x) - p(x)\} \sim p^{(2)}(x)(\beta^2/3 + c^2)/2,$$

среднеквадратическое отклонение

$$W_2(\bar{p}(x)) = M\left\{\left\|\bar{p}(x) - p(x)\right\|^2\right\} \sim \left\|\Phi(u)\right\|^2 \Big/ nc + \left\|p^{(2)}(x)\right\|^2 (\beta^2/3 + c^2)^2/4 +$$

$$+\left(\left\|\Phi(u)\right\|^2 \beta/c\right)\left(2\left\|p(x)\right\|^2 + \beta^4\left\|p^{(2)}(x)\right\|^2/18\right) + \beta c\left\|u\Phi(u)\right\|^2 \times$$

$$\times\left(2\left\|p^{(1)}(x)\right\|^2 + \beta^2\left\|p^{(2)}(x)\right\|^2/3\right) + \beta c^3\left\|p^{(2)}(x)\right\|^2\left\|u^2\Phi(u)\right\|^2/2.$$

Сравнение асимптотических свойств $\bar{p}(x)$ и непараметрической оценки $\bar{p}_1(x)$ типа Парзена [2], восстанавливаемой по выборке объема N, позволяет получить оценку сверху разности среднеквадратических отклонений

$$W_2(\bar{p}(x)) - W_2(\bar{p}_1(x)) < \frac{\left\|\Phi(u)\right\|^2}{c}\left(\frac{1}{n} - \frac{1}{N} + 2\beta\right).$$

Смещение $\bar{p}(x)$ несколько выше, чем смещение традиционной оценки Парзена.

Преимущество регрессионной плотности вероятности $\bar{p}(x)$ заключается не только в повышении вычислительной эффективности непараметрических алгоритмов, синтезируемых на ее основе, но и в упрощении задачи оптимизации $\bar{p}(x)$ по параметрам c в режиме «скользящего экзамена» по выборке $(z^i,\ P^i/2\beta,\ i=\overline{1,N})$.

В многомерном случае при $x = (x_1, \ldots, x_k)$ регрессионная оценка плотности вероятности запишется в виде

$$\bar{p}(x) = \left(\prod_{\nu=1}^{k} c_\nu\right)^{-1} \sum_{i=1}^{N^k} P^i \prod_{\nu=1}^{k} \Phi\left(\frac{x_\nu - x_\nu^i}{c_\nu}\right). \qquad (3)$$

На основе предлагаемых оценок плотности вероятности (1), (3) возможна реализация последовательной вычислительной процедуры

$$\bar{A}_s(\bar{p}_s(x), r_{s+1}(x)), \quad s=\overline{1,m}, \qquad (4)$$

каждый последующий $(s+1)$-й этап которой реализуется непараметрическим алгоритмом $\overline{A}_s(\cdot)$ в некоторой окрестности предыдущего решения $r_s(x)$ и уточняет его $r_{s+1}(x)$. Причем синтез алгоритмов $\overline{A}_s(\cdot)$ осуществляется на основе непараметрических оценок плотности вероятности $\overline{p}_s(x)$, восстанавливаемых по выборкам ограниченного объема N_s, значительно меньшего исходного n.

Установлено, что при $N_s = N$, $s = \overline{1, m}$, с увеличением числа этапов s процедуры последовательной обработки информации среднеквадратическая ошибка оценивания $p(x)$ снижается пропорционально значению $(N/2)^{4(s-1)/5}$. Особенность структуры регрессионных оценок плотности вероятности позволяет решить проблему их доверительного оценивания по следующей методике:

– рассчитываются доверительные границы P_1^i, P_2^i для частоты P^i попадания случайной величины x в i-й интервал при некотором значении коэффициента доверия α;

– формируются статистические выборки $(z^i, P_1^i, i = \overline{1, N}), (z^i, P_2^i, i = \overline{1, N})$;

– строятся верхняя $\tilde{p}_1(x)$ и нижняя $\tilde{p}_2(x)$ толерантные границы регрессионной оценки плотности вероятности

$$\tilde{p}_\nu(x) = c^{-1} \sum_{j=1}^{N} P_\nu^j \Phi\left(\frac{x - z^j}{c}\right), \quad \nu = 1, 2.$$

По аналогии осуществляется доверительное оценивание других непараметрических решающих функций.

Применение метода декомпозиции обучающей выборки в задаче распознавания образов. Синтез непараметрических алгоритмов распознавания образов основан на оценивании байесовского уравнения разделяющей поверхности [3]. В случае двух классов $\Omega_1(x)$ и $\Omega_2(x)$ уравнение разделяющей поверхности имеет вид

$$f_{12}(x) = p_2(x) - p_1(x) \tag{5}$$

и может быть оценено с помощью непараметрических методов статистики по обучающей выборке $V = (x^i, \sigma(x^i), i = \overline{1, n})$.

Примем в качестве оценок плотностей вероятности $x = (x_1, \ldots, x_k)$ в классах $\Omega_1(x)$ и $\Omega_2(x)$ процедуры типа (3). Тогда непараметрическая оценка $f_{12}(x)$ (5) представляется выражением

$$\bar{f}_{12}(x) = \left(\prod_{\nu=1}^{k} c_\nu\right)^{-1} \sum_{i=1}^{N^k} \sigma_{12}(x^i) P^i \prod_{\nu=1}^{k} \Phi\left(\frac{x_\nu - x_\nu^i}{c_\nu}\right), \tag{6}$$

где

$$\sigma_{12}(x^i) = \begin{cases} -1, & \text{если } x^i \in \Omega_1(x), \\ 1, & \text{если } x^i \in \Omega_2(x); \end{cases}$$

$N^k = N_j^k + N_{\bar{j}}^k$ (N_j^k, $N_{\bar{j}}^k$ – количество элементов выборки, на основании которых формируются оценки плотностей $\bar{p}_1(x), \bar{p}_2(x)$).

Асимптотическую сходимость $\bar{f}_{12}(x)$ нетрудно доказать, используя утверждения теоремы 1.

Оптимизация непараметрической оценки решающей функции (6) по параметрам β_ν, c_ν, $\nu = \overline{1,k}$, осуществляется из условия минимума статистической оценки вероятности ошибки распознавания образов, формируемой по контрольной выборке $V_1 \subset V$. Проводя предварительную нормировку компонент x_ν, $\nu = \overline{1,k}$, относительно их среднего значения, размерность задачи оптимизации можно сократить до двух параметров: $\beta_\nu = \beta, c_\nu = c, \ \nu = \overline{1,k}$.

По результатам статистического моделирования применение предлагаемых алгоритмов при снижении временных затрат на порядок обеспечивает увеличение ошибки распознавания образов не более чем на 5 % по сравнению с прямыми методами обработки. При случайной стратегии формирования выборки меньшего объема $N^k < n$ отмечается значительное увеличение ошибки распознавания (до 15 %).

Синтез и анализ непараметрической регрессии на основе метода декомпозиции выборки. Пусть переменные регрессии $y = \varphi(x)$ скаляры x, $y \in R^1$. В соответствии с методикой синтеза $\bar{p}(x)$ (1) нетрудно получить выражения для оценки плотности вероятности двумерной случайной величины (x, y):

$$\bar{p}(x,y) = c^{-2} \sum_{i=1}^{N^2} P^i \Phi\left(\frac{x - x^i}{c}\right) \Phi\left(\frac{y - y^i}{c}\right),$$

где объем N^2 рабочей выборки $(x^i, y^i, P^i, i = \overline{1, N^2})$, сформированной на основе обучающей $V = (x^j, y^j, j = \overline{1,n})$, значительно меньше n.

Тогда с учетом $p(y/x) = 2N\beta p(x,y)$ непараметрическая оценка регрессии представляется в виде

$$\bar{y} = \varphi(x) = \frac{2N\beta}{c} \sum_{i=1}^{N^2} y^i P^i \Phi\left(\frac{x - x^i}{c}\right). \qquad (7)$$

Справедлива следующая

Теорема 2. Пусть

1) функции $\varphi(x)$, $p(x,y)$, $p(x)$ ограничены и непрерывны со всеми своими производными до второго порядка включительно;

2) ядерные функции $\Phi(u) \in H$;

3) последовательности $c = c(n) \to 0$, $\beta = \beta(n) \to 0$ при $n \to \infty$, а $nc \to \infty$.

Тогда смещение

$$M(\bar{\varphi}(x) - \varphi(x)) \sim \varphi(x)\left(\frac{p(x)}{p_1(x)} - 1\right) + \left(\frac{\beta^2}{3} + c^2\right) \times$$

$$\times\left(\frac{\varphi(x)p^{(2)}(x)}{2p_1(x)}+\frac{\varphi^{(1)}(x)p^{(1)}(x)}{p_1(x)}+\varphi^{(2)}(x)p(x)\right)+$$

$$+\frac{\varphi^{(2)}(x)p^{(2)}(x)}{6}\beta^2+O(c^4,\ \beta^4,\ c^2\beta^2),$$

квадратическое отклонение

$$M(\overline{\varphi}(x)-\varphi(x))^2\sim\varphi^2(x)(p(x)/p_1(x)-1)^2+(\beta^2/3+c^2)\times$$

$$\times[(p(x)/p_1(x)-1)(\varphi^2(x)p^{(2)}(x)+2\varphi(x)\varphi^{(1)}(x)p^{(1)}(x))/p_1(x)+$$

$$+2\varphi(x)p(x)\varphi^{(2)}(x)(p(x)/2p_1^2(x)-1)]+\beta^2(\varphi(x)p^{(1)}(x)+p(x)\varphi^{(1)}(x)+$$

$$+p^2(x))/(3p_1^2(x))+c^2\varphi^2(x)p^2(x)/4+$$

$$+(\varphi^2(x)p(x))/(ncp_1^2(x))\int\Phi^2(u)du+O(c^4,\ \beta^4,\ c^2\beta^2,\ \beta^2/nc,\ c/n).$$

Таким образом, статистика (7) обладает свойством асимптотической несмещенности и сходимостью в квадратическом, если закон распределения исходной выборки наблюдений x равномерный, т. е. $p(x)=p_1(x)$, что возможно при проведении активного эксперимента с исследуемой зависимостью $y=\varphi(x)$.

Для устранения смещения умножим $\bar{y}=\overline{\varphi}(x)$ на $p_1(x)/\overline{p}(x)$, получим

$$\bar{\bar{y}}=\sum_{i=1}^{N^2}y^i\alpha^i(x)\Big/\sum_{i=1}^{N^2}\alpha^i(x), \tag{8}$$

где $\alpha^i(x)=P^i\Phi\left(\frac{x-x^i}{c}\right)$.

В многомерном случае при $x=(x_1,\ldots,x_k)$ структура оценки (8) не меняется, а весовая функция принимает вид

$$\alpha^i(x)=P^i\prod_{\nu=1}^{k}\Phi\left(\frac{x_\nu-x_\nu^i}{c_\nu}\right),\quad i=\overline{1,N^k}.$$

Анализ множеств случайных величин при восстановлении стохастических зависимостей. Определим на элементах множеств X^i, Y^i исходной обучающей выборки $(x^i,\ y^i,\ i=\overline{1,N})$ непараметрические оценки плотностей вероятности [4]:

$$\overline{p}_i(x)\ \forall\ x\in X^i,\qquad \overline{p}_i(y)\ \forall\ y\in Y^i,\qquad i=\overline{1,N}.$$

Тогда два множества X^i, X^j близки, если соответствующие им статистические оценки функций распределения $\overline{F}_i(x)$, $\overline{F}_j(x)$ тождественны, т. е. справедлива гипотеза

$$H_0\colon\ \overline{F}_i(x) \equiv \overline{F}_j(x) \tag{9}$$

с некоторым уровнем доверия β.

Пусть вероятностные законы распределения элементов случайных множеств X^i, X^j оцениваются по выборкам объема N_i, N_j. В качестве критерия близости между оценками функций распределения $\overline{F}_i(x)$ и $\overline{F}_j(x)$ может быть взята ядерная мера

$$h(X^j, X^i) = \prod_{\nu=1}^{k} \Phi\left(\overline{F}_i(x_\nu), \overline{F}_j(x_\nu)\right);$$

$$\Phi\left(\overline{F}_i(x_\nu), \overline{F}_j(x_\nu)\right) = \begin{cases} (D_\beta + c)^{-1}, & \text{если } \Delta \le D_\beta, \\ \dfrac{D_\beta + c - \Delta}{D_\beta + c}, & \text{если } D_\beta \le \Delta \le D_\beta + c, \\ 0, & \text{если } \Delta > D_\beta; \end{cases} \tag{10}$$

$$\Delta = \max_{x_\nu} |\overline{F}_i(x) - \overline{F}_j(x)|.$$

Предложенная мера близости (10) основана на использовании критерия Смирнова для проверки гипотезы H_0 (9) с уровнем доверия β:

$$D_\beta = \sqrt{-\ln\frac{\beta}{2}\left(\frac{1}{N_i} + \frac{1}{N_j}\right)\Big/2}.$$

Превышение порогового значения D_β критерия Смирнова означает нарушение гипотезы H_0.

На основе введенных понятий запишем непараметрическую модель преобразования случайных множеств

$$\overline{\overline{p}}(y) = \frac{\sum_{i=1}^{N} \overline{p}_i(y)\, h(X, X^i)}{\sum_{i=1}^{N} h(X, X^i)}, \tag{11}$$

$$G\colon\ \overline{\overline{p}}(y) \to \overline{Y}.$$

Оператор G является датчиком случайных величин, с помощью которого формируется оценка $\overline{Y}$ множества Y, соответствующего X.

Выбор порогового значения c осуществляется из условия минимума ошибки оценивания $\overline{p}(y)$ решающим правилом (11) в режиме «скользящего экзамена».

Если плотность $p(y)$ представима в виде линейного функционала от $p_j(y)$, $j=\overline{1, N_1}$, $N_1 < N$:

$$p(y)=\frac{1}{N_1}\sum_{j=1}^{N_1} p_j(y), \tag{12}$$

то при ограниченных $p_j^{(2)}(y)$, $j=\overline{1, N_1}$, и $c_j^2(N_j)\to 0$ при $N_j \to \infty$, $j=\overline{1, N_1}$, непараметрическая статистика $\bar{\bar{p}}(y)$ является несмещенной оценкой $p(y)$:

$$M(\bar{\bar{p}}(y)-p(y))\sim \frac{1}{2N_1}\sum_{j=1}^{N_1} c_j^2 p_j^{(2)}(y)+O(c^4).$$

При дополнительных условиях $N_j c_j \to \infty \; \forall \, N_j \to \infty$, $c_j \to 0, j=\overline{1, N_1}$, имеет место асимптотическая сходимость $\bar{\bar{p}}(y)$ к $p(y)$ в среднеквадратическом

$$M\int(\bar{\bar{p}}(y)-p(y))^2\,dy \sim \frac{1}{N_1^2}\sum_{j=1}^{N_1}\left[\frac{\|\Phi(u)\|^2}{N_j c_j}+\frac{c_j^4\left\|p_j^{(2)}(y)\right\|^2}{4}(1+N_1)\right].$$

Заключение. Рассмотрены непараметрические модели анализа множеств случайных величин в задачах восстановления стохастических зависимостей и распознавания образов. Идея предлагаемого подхода заключается в замене операций над множествами операциями с вероятностными законами распределения их элементов либо с параметрами законов с использованием методов локальной аппроксимации. Полученные результаты позволяют с новых теоретических позиций решить проблемы доверительного оценивания непараметрических статистик и их оптимизации. Наиболее перспективным направлением применения разработанных непараметрических моделей является обработка больших массивов статистических данных, часто встречающихся при исследовании медико-биологических и экологических систем.

СПИСОК ЛИТЕРАТУРЫ

1. **Епаненчиков В. А.** Непараметрическая оценка многомерной плотности вероятности // Теория вероятности и ее применения. 1969. **14**, вып. 1. С. 156.
2. **Parzen E.** On estimation of a probability density function and mode // Ann. Math. Stat. 1962. **33**. P. 1065.
3. **Цыпкин Я. З.** Основы теории обучающихся систем. М.: Наука, 1970.
4. **Лапко А. В., Лапко В. А., Соколов М. И., Ченцов С. В.** Непараметрические системы классификации. Новосибирск: Наука, 2000.

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

Поступила в редакцию
24 апреля 2002 г.