Стохастические системы
2013. №1(35)

УДК 519.7

© 2013 г. **А.В. Лапко**, д-р техн. наук,
В.А. Лапко, д-р техн. наук
(Институт вычислительного моделирования СО РАН, Красноярск)
(Сибирский государственный аэрокосмический университет имени академика М.Ф. Решетнева, Красноярск)

АНАЛИЗ СВОЙСТВ НЕПАРАМЕТРИЧЕСКИХ ОЦЕНОК СМЕСИ ПЛОТНОСТЕЙ ВЕРОЯТНОСТИ ПРИ РАЗЛИЧНЫХ УСЛОВИЯХ РАСПРЕДЕЛЕНИЯ СТАТИСТИЧЕСКИХ ДАННЫХ

Рассматриваются непараметрические оценки смеси плотностей вероятности типа Розенблатта – Парзена, синтез которых основан на декомпозиции исходных статистических данных по объему. Исследуется зависимость их свойств от количества составляющих смеси и уровня неравномерности распределения наблюдений между ними.
Ключевые слова: плотность вероятности, непараметрическая оценка, декомпозиция статистических данных, асимптотические свойства.

Введение

Использование непараметрических методов обработки информации, основанных на оценках плотности вероятности типа Розенблатта – Парзена, является одним из активно развивающихся направлений теории принятия решений при априорной неопределенности [1 – 5]. Однако по мере усложнения изучаемых объектов появляются методологические и вычислительные трудности применения традиционных непараметрических алгоритмов, особенно при анализе неоднородных данных большого объема [6, 7].

Перспективное направление «обхода» возникающих проблем состоит в применении принципов декомпозиции систем и методов коллективного оценивания. Для его реализации в работах [8, 9] предложена непараметрическая оценка смеси плотностей вероятности, использование которой при синтезе алгоритмов обработки информации позволило значительно повысить их вычислительную эффективность в условиях больших выборок. Идея предлагаемого подхода состоит в декомпозиции исходных статистических данных по их объему на равные группы наблюдений одномерной случайной величины, построении на этой основе семейства непараметрических оценок плотности вероятности и последующем синтезе их смеси в виде линейного коллектива. Полученные результаты обобще-

ны на многомерный случай [10, 11], развиты при построении непараметрических решающих функций в задаче распознавания образов [12] и восстановлении стохастических зависимостей [13, 14].

Цель данной работы состоит в исследовании зависимости аппроксимационных свойств непараметрической оценки смеси плотностей вероятности от уровня неравномерности распределения исходных статистических данных между ее составляющими и их количеством.

Непараметрическая оценка смеси плотностей вероятности

Имеется выборка $V = \left(x^i, i = \overline{1, n}\right)$ из n статистически независимых наблюдений одномерной случайной величины x с априори неизвестной плотностью вероятности $p(x)$. Выборка V допускает разбиение на N групп наблюдений $V_j = \left(x^i, i \in I_j\right)$, $j = \overline{1, N}$. Множество номеров наблюдений x в группе под номером j обозначается через I_j, а количество элементов данного множества – n_j.

По каждой выборке V_j построим непараметрическую оценку плотности вероятности случайной величины x [15]:

$$\overline{p}_j(x) = \frac{1}{n_j c_j} \sum_{i \in I_j} \Phi\left(\frac{x - x^i}{c_j}\right), \quad j = \overline{1, N}, \tag{1}$$

где ядерные функции $\Phi(\cdot)$ удовлетворяют условиям *H*:

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

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

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

Здесь и далее бесконечные пределы интегрирования опускаются. Параметры $c_j = c_j(n_j)$ ядерных функций убывают с ростом n_j, $j = \overline{1, N}$.

В качестве приближения $p(x)$ по статистическим данным примем смесь непараметрических оценок плотностей вероятности типа (1):

$$\overline{\overline{p}}(x) = \sum_{j=1}^{N} \frac{n_j}{n} \overline{p}_j(x). \tag{2}$$

Статистика (2) учитывает не только особенности условий контроля значений случайной величины x, но и допускает использование технологии параллельных вычислений при оценивании плотностей вероятностей в условиях больших выборок.

Асимптотические свойства $\overline{\overline{p}}(x)$ определяются следующим утверждением.

Теорема. Пусть плотность вероятности $p(x)$ одномерной случайной величины и первые две ее производные по x ограниченны и непрерывны; ядерные функции $\Phi(u)$ удовлетворяют условиям *H*; последовательности $c_j = c(n_j)$ коэффициентов размытости ядерных функций таковы, что при $n \to \infty$ значения

$c_j \to 0$, а $nc_j \to \infty$, $j = \overline{1, N}$.

Тогда при конечных значениях N непараметрическая оценка плотности вероятности $\overline{\overline{p}}(x)$ смеси плотностей вероятности обладает свойством асимптотической несмещенности и состоятельности.

Доказательство.

1. По определению:

$$M\left(\overline{\overline{p}}(x)\right) = \frac{1}{n}\sum_{j=1}^{N}\frac{1}{c_j}\sum_{i\in I_j} M\left(\Phi\left(\frac{x - x^i}{c_j}\right)\right) =$$

$$= \frac{1}{n}\sum_{j=1}^{N}\frac{1}{c_j}\sum_{i\in I_j}\int\Phi\left(\frac{x - x^i}{c_j}\right)p\left(x^i\right)dx^i = \frac{1}{n}\sum_{j=1}^{N}\frac{n_j}{c_j}\int\Phi\left(\frac{x - t}{c_j}\right)p(t)\,dt,$$

где M – знак математического ожидания.

При выполнении этих преобразований учитывается, что элементы статистических выборок V_j, $j = \overline{1, N}$ являются значениями одной и той же случайной величины t с плотностью вероятности $p(t)$.

Проведем замену переменных $x - t = c_j u_j$ и, разлагая функции $p\left(x - c_j u_j\right)$, $j = \overline{1, N}$ в ряд Тейлора в точке x, с учетом свойств ядерных функций при достаточно больших n, получим асимптотическое выражение

$$M\left(\overline{\overline{p}}(x)\right) \sim p(x) + \frac{p^{(2)}(x)}{2n}\sum_{j=1}^{N} n_j\, c_j^2, \qquad (3)$$

где $p^{(2)}(x)$ – вторая производная плотности вероятности $p(x)$ по x.

Так как отношение $\left(n_j / n\right) < 1$, то при конечных N из условия $c_j \to 0$ при $n_j \to \infty$ следует свойство асимптотической несмещенности статистики (2).

2. Для доказательства сходимости в среднеквадратическом рассмотрим выражение

$$M\left(\overline{\overline{p}}(x) - p(x)\right)^2 = M\left(\sum_{j=1}^{N}\frac{n_j}{n}\left(p(x) - \overline{p}_j(x)\right)\right)^2 = \sum_{j=1}^{N}\frac{n_j^2}{n^2}M\left(p(x) - \overline{p}_j(x)\right)^2 +$$

$$+ \sum_{j=1}^{N}\sum_{\substack{t=1\\ t\neq j}}^{N}\frac{n_j n_t}{n^2}M\left(\left(p(x) - \overline{p}_j(x)\right)\left(p(x) - \overline{p}_t(x)\right)\right). \qquad (4)$$

С учетом статистической независимости элементов выборок V_j и V_t проведем анализ второй части выражения (4)

$$M\left(\left(p(x) - \overline{p}_j(x)\right)\left(p(x) - \overline{p}_t(x)\right)\right) =$$

$$= p^2(x) - p(x)\,M\left(\overline{p}_t(x)\right) - p(x)\,M\left(\overline{p}_j(x)\right) + M\left(\overline{p}_t(x)\right)M\left(\overline{p}_j(x)\right). \qquad (5)$$

Известно, что асимптотические выражения [9, 16]

$$M\left(\overline{p}_j(x)\right)\sim p(x)+\frac{c_j^2}{2}p^{(2)}(x),\ M\left(\overline{p}_t(x)\right)\sim p(x)+\frac{c_t^2}{2}p^{(2)}(x).$$

Тогда при достаточно больших значениях n_j, $j=\overline{1,N}$ выражение (5) преобразуется к виду

$$M\left(\left(p(x)-\overline{p}_j(x)\right)\left(p(x)-\overline{p}_t(x)\right)\right)\sim\frac{c_j^2\,c_t^2}{4}\left(p^{(2)}(x)\right)^2. \qquad (6)$$

В работах [9, 16] получено и используется асимптотическое выражение для среднеквадратического отклонения

$$M\left(\overline{p}(x)-p(x)\right)^2\sim\frac{1}{n_j\,c_j}\int\Phi^2(u)du+\frac{c_j^4}{4}\left(p^{(2)}(x)\right)^2, \qquad (7)$$

непараметрической оценки $\overline{p}_j(x)$ плотности вероятности от $p(x)$, составляющих первую часть выражения (4).

С учетом (6), (7), выражение (4) представляется как

$$\begin{aligned}M\left(\overline{\overline{p}}(x)-p(x)\right)^2\sim&\frac{1}{n^2}\int\Phi^2(u)du\sum_{j=1}^{N}\frac{n_j}{c_j}+\frac{1}{4n^2}\left(p^{(2)}(x)\right)^2\sum_{j=1}^{N}n_j^2\,c_j^4+\\&+\frac{1}{4n^2}\left(p^{(2)}(x)\right)^2\sum_{j=1}^{N}\sum_{\substack{t=1\\t\neq j}}^{N}n_j\,c_j^2\,n_t\,c_t^2=\frac{1}{n^2}\int\Phi^2(u)du\sum_{j=1}^{N}\frac{n_j}{c_j}+\\&+\frac{1}{4n^2}\left(p^{(2)}(x)\right)^2\left(\sum_{j=1}^{N}n_j\,c_j^2\right)^2.\end{aligned} \qquad (8)$$

Нетрудно заметить, что с учетом $\frac{n_j}{n}<1$ в условиях $c_j\to 0$, $nc_j\to\infty$ при $n_j\to\infty$, $j=\overline{1,N}$ оценка смеси плотностей вероятности (2) сходится в среднеквадратическом к $p(x)$, а с учетом свойства ее асимптотической несмещенности (3) является состоятельной.

При $N=1$ выражение (8) соответствует результату, полученному в работе [16], а при равных объемах $n_j=n/N$ групп V_j, $j=\overline{1,N}$ наблюдений случайной величины x совпадает с выводом работы [9], что подтверждает корректность выполненных преобразований.

Анализ аппроксимационных свойств статистики $\overline{\overline{p}}(x)$

Исследуем отношение минимальных значений W_2, $\overline{W}_2$ среднеквадратического отклонения $\int M\left(p(x)-\overline{\overline{p}}(x)\right)^2dx$, соответствующих неравномерному и равномерному распределению объема n исходных данных V между группами наблюдений V_j, $j=\overline{1,N}$ при оптимальных значениях параметров c_j, $j=\overline{1,N}$ ста-

тистики (2).

Известно, что оптимальное значение $\overline{c}_j$ коэффициента размытости ядерных функций непараметрической статистики типа $\overline{p}_j(x)$ (1), минимизирующего асимптотическое выражение

$$\frac{\left\|\Phi(u)\right\|^2}{n_j\, c_j} + \frac{c_j^4}{4}\left\|p^{(2)}(x)\right\|^2$$

среднеквадратического отклонения $\int M\left(p(x) - \overline{\overline{p}}(x)\right)^2 dx,$ определяется формулой

$$\overline{c}_j = \left(\frac{\left\|\Phi(u)\right\|^2}{n_j \left\|p^{(2)}(x)\right\|^2}\right)^{\frac{1}{5}}. \qquad (9)$$

Здесь приняты следующие обозначения

$\left\|\Phi(u)\right\|^2 = \int \Phi^2(u)du$, $\left\|p^{(2)}(x)\right\|^2 = \int\left(p^{(2)}(x)\right)^2 dx$.

Минимальное значение

$$\overline{W}_2 = \left(\left(\left\|\Phi(u)\right\|^2\right)^4 \left\|p^{(2)}(x)\right\|^2\right)^{\frac{1}{5}} \frac{4+N}{4\left(N\, n^4\right)^{1/5}} \qquad (10)$$

асимптотического выражения среднеквадратического отклонения $\overline{\overline{p}}(x)$ при равномерном распределении исходных наблюдений V между их группами V_j, $j = \overline{1, N}$ нетрудно получить путем преобразования интеграла от выражения (8) по x при $n_j = n/N$ и

$$\overline{c} = \overline{c}_j = \left(\frac{N\left\|\Phi(u)\right\|^2}{n\left\|p^{(2)}(x)\right\|^2}\right)^{\frac{1}{5}},\ j = \overline{1, N}.$$

Для формирования условий неравномерного распределения исходных данных между группами наблюдений V_j, $j = \overline{1, N}$ разобьем их на два набора $V^1(m) = \left(V_j,\ j = \overline{1, m}\right)$ и $V^2(m) = \left(V_j,\ j = \overline{m+1, N}\right)$. Объем наблюдений в каждой группе из первого набора одинаков $n^1 = n_j = \frac{n}{2N}$, $j = \overline{1, m}$. Во втором наборе оставшиеся наблюдения в количестве $n(2) = n - \frac{nm}{2N}$ распределяются между группами равномерно при $n^2 = n_j = \frac{n(2)}{(N-m)}$, $j = \overline{m+1, N}$. Косвенным показателем

уровня неравномерности распределения наблюдений между наборами $V^1(m)$, $V^2(m)$ является значение m.

С учетом выражения (8) вычислим минимальное значение W_2 среднеквадратического отклонения

$$\int M\left(p(x)-\bar{\bar{p}}(x)\right)^2 dx \qquad (11)$$

в принятых условиях неравномерности распределения исходных данных.

Заметим, что составляющие смеси типа (1), восстанавливаемые по статистическим данным V_j, $j=\overline{1,m}$, характеризуются оптимальным коэффициентом размытости $\bar{c}^1 = 2^{1/5}\,\bar{c}$. При этом непараметрические оценки плотности вероятности $\bar{p}_j(x)$, восстанавливаемые по группам наблюдений V_j, $j=\overline{m+1,N}$, имеют оптимальный коэффициент размытости

$$\bar{c}^2 = \bar{c}\left(\frac{2(N-m)}{2N-m}\right)^{\frac{1}{5}}.$$

Тогда асимптотическое выражение среднеквадратического отклонения (11) при оптимальных значениях $\bar{c}^1$, $\bar{c}^2$ запишется в виде

$$W_2 = \left(\frac{\left(\left\|\Phi(u)\right\|^2\right)^4 \left\|p^{(2)}(x)\right\|^2}{(2N)^6 n^4}\right)^{\frac{1}{5}} F(m,n), \qquad (12)$$

где $F(m,n) = \left[m + \left(\frac{(2N-m)^6}{N-m}\right)^{\frac{1}{5}} + \frac{1}{4}\left(m + \left((2N-m)^3(N-m)^2\right)^{1/5}\right)^2\right]$.

При $N=1$ и $m=0$ из выражения (12) следует результат работы [16], а при $m=0$ – вывод работы [9].

Проведем анализ отношения

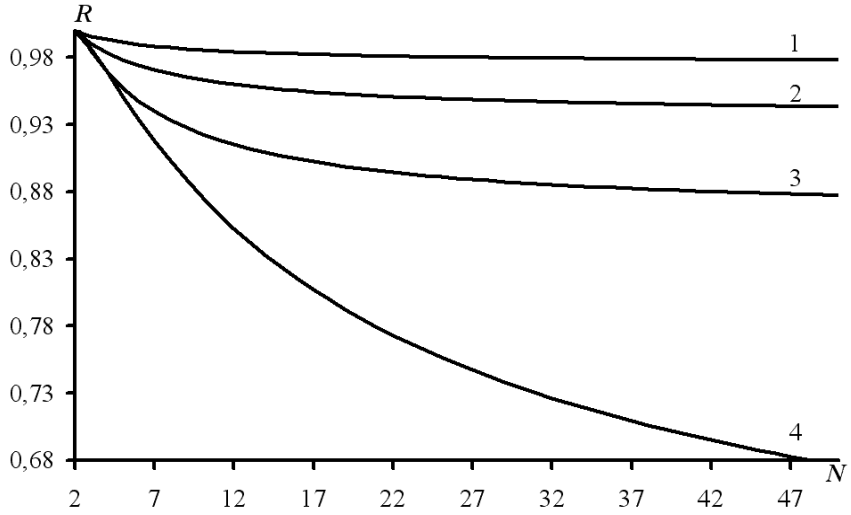
$$R = \frac{W_2}{\overline{W}_2} = \frac{2^{4/5}}{N(4+N)} F(m,n)$$

при конкретных значениях $m = 0.25N,\ 0.5N,\ 0.75N,\ N-1$.

С ростом уровня неравномерности распределения исходных данных между составляющими оценки смеси $\bar{\bar{p}}(x)$ наблюдается уменьшение значений отношения $R<1$ (рисунок). Отмеченное повышение аппроксимационных свойств $\bar{\bar{p}}(x)$ по сравнению с равномерным характером распределения статистической информации объясняется наличием в наборе V_j, $j=\overline{m+1,N}$ выборок V_j с объемом n_j больше либо значительно больше n/N. Например, если $m=N-1$ объем $n(N+1)/2N$ выборки V_N в $(N+1)/2$ больше количества $\bar{n}=n/N$ наблюдений,

используемых при оценивании составляющих смеси с равномерным распределением статистических данных. При $m = N/2$ значение $n_j = 3n/(2N)$, $j = \overline{m+1, N}$ в 1.5 раза больше по сравнению с $\overline{n}$.

Причем их значимость в формировании свойств оценки смеси (2) возрастает пропорционально значению n_j/n, $j = \overline{m+1, N}$. Так, при $m = N-1$ вес статистики $\overline{p}_N(x)$ в оценке (2) определяется значением $(N+1)/2N$. Кривым 1, 2, 3, 4, (см. рисунок), соответствуют значения $m = 0.25N, 0.5N, 0.75N, N-1$.



Зависимость отношения R среднеквадратических отклонений непараметрических оценок смеси плотностей вероятности при неравномерном и равномерном распределении исходных статистических данных между ее составляющими от их количества *N*.

Увеличение объема статистических данных V_j, $j = \overline{m+1, N}$ способствует снижению смещения и дисперсии составляющих смеси. С уменьшением значения m объем статистических данных при восстановлении составляющих смеси $\overline{\overline{p}}(x)$ становится сопоставимым со значением n/N. Поэтому аппроксимационные свойства исследуемой статистики становятся сопоставимыми со свойствами смеси с равномерным распределением исходных данных между ее составляющими, что проявляется в стремлении отношения R к единице при уменьшении значений m.

Существует пороговое значение $N \approx 20$ количества составляющих оценки смеси $\overline{\overline{p}}(x)$, превышение которого не оказывает существенного влияния на ее аппроксимационные свойства, особенно при малых значениях m.

Заключение

Непараметрическая оценка смеси плотностей вероятности при неравномерном распределении исходных статистических данным между ее составляющими обладает свойствами асимптотической несмещенности и состоятельности. Асимптотическое выражение ее среднеквадратического отклонения обобщает результаты, полученные при исследовании традиционной оценки плотности вероятности типа Розенблатта – Парзена и с равномерным характером распределения статистических данных между ее составляющими. Аппроксимационные свойства исследуемой статистики выше по сравнению с оценкой смеси, синтез которой ос-

нован на равномерном характере распределения исходных данных. Установлено пороговое значение количества составляющих смеси, превышение которого не оказывает существенного влияния на изменения ее асимптотического выражения среднеквадратического отклонения.

ЛИТЕРАТУРА

1. *Лапко В.А., Капустин А.Н.* Синтез нелинейных непараметрических коллективов решающих правил в задачах распознавания образов // Автометрия. – 2006. – №6. – С.26-33.
2. *Лапко А.В., Лапко В.А*. Непараметрические алгоритмы распознавания образов при случайных значениях коэффициентов размытости ядерных функций // Автометрия. – 2007. – № 5. – С.47-55.
3. *Лапко А.В., Лапко В.А., Ярославцев С.Г.* Разработка и исследование гибридных алгоритмов в задачах распознавания образов // Автометрия. – 2006. – №1. – С.32-39.
4. *Лапко А.В., Лапко В.А*. Гибридные модели стохастических зависимостей // Автометрия. – 2002. – №5. – С.38-48.
5. *Лапко А.В., Лапко В.А.* Анализ непараметрических алгоритмов распознавания образов в условиях пропуска данных // Автометрия. – 2008. – № 3. – С. 65-74.
6. *Лапко А.В., Лапко В.А.* Разработка и исследование двухуровневых непараметрических систем классификации // Автометрия. – 2010. – № 1. – С.70-78.
7. *Лапко А.В., Лапко В.А.* Синтез структуры семейства непараметрических решающих функций в задаче распознавания образов // Автометрия. – 2011. – № 4. – С.76-82.
8. *Лапко А.В., Лапко В.А., Егорочкин И.А.* Непараметрические оценки смеси плотностей вероятности и их применение в задаче распознавания образов // Системы управления и информационные технологии. – 2009. – 1(35). – С.60-64.
9. *Лапко В.А., Варочкин С.С., Егорочкин И.А*. Разработка и исследование непараметрической оценки плотности вероятности, основанной на принципе декомпозиции обучающей выборки по ее объему // Вестник СибГАУ. – 2009. – №1(22). – Ч.1. – С.45-49.
10. *Лапко А.В., Лапко В.А.* Анализ свойств смеси непараметрических оценок плотности вероятности многомерной случайной величины // Вестник СибГАУ. – 2010. – №2(28). – С.32-35.
11. *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.
12. *Лапко А.В., Лапко В.А.* Непараметрическая оценка уравнения разделяющей поверхности в условиях больших выборок и ее свойства // Системы управления и информационные технологии. – 2010. – №1.2 (39). – С.300-304.
13. *Лапко А.В., Лапко В.А., Варочкин С.С.* Коллектив непараметрических регрессий, основанный на принципе декомпозиции обучающей выборки // Вестник СибГАУ. – 2009. – №1(22). – Ч.2. – С. 38-40.
14. Лапко А.В., Лапко В.А. Коллектив многомерных непараметрических регрессий, основанный на декомпозиции обучающей выборки по ее объему // Вестник СибГАУ. – 2012. – №3 (43). – С.42-46.
15. *Parzen E.* On estimation of a probability density function and mode // Ann. Math. Statistic. – 1962. – Vol. 33. – P.1065-1076.
16. *Епанечников В.А.* Непараметрическая оценка многомерной плотности вероятности // Теория вероятности и ее применения. – 1969. – Т.14, №1. – С.156-161.

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