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)

ДИСКРЕТИЗАЦИЯ ИНТЕРВАЛА ИЗМЕНЕНИЯ ЗНАЧЕНИЙ СЛУЧАЙНОЙ ВЕЛИЧИНЫ НА ОСНОВЕ РЕЗУЛЬТАТОВ ОПТИМИЗАЦИИ НЕПАРАМЕТРИЧЕСКОЙ ОЦЕНКИ ПЛОТНОСТИ ВЕРОЯТНОСТИ

Предлагается новая методика дискретизации интервала изменения значений случайной величины, основанная на оптимизации непараметрической оценки плотности вероятности типа Розенблатта – Парзена. Приводятся результаты ее сравнения с традиционными процедурами дискретизации.
Ключевые слова: метод дискретизации, непараметрическая оценка, плотность вероятности, оптимизация, коэффициент размытости.

Введение

Пусть имеется выборка $V=\left(x^i,\, i=\overline{1,\, n}\right)$, состоящая из n независимых наблюдений одномерной случайной величины с неизвестной плотностью вероятности $p(x)$, которая непрерывна и ограничена со всеми своими производными хотя бы до второго порядка включительно. В качестве приближения по эмпирическим данным искомой плотности $p(x)$ примем ее непараметрическую оценку [1]

$$\overline{p}(x)=\frac{1}{nc}\sum_{i=1}^{n}\Phi\left(\frac{x-x^i}{c}\right), \qquad (1)$$

где ядерные функции $\Phi(u)$ обладают следующими свойствами:

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

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

$$\int u^m\,\Phi(u)du<\infty \text{ при } 0\le m<\infty.$$

Здесь и далее бесконечные пределы интегрирования опускаются.

Коэффициенты размытости $c=c(n)$ ядерных функций непараметрической оценки (1) убывают с увеличением объема n выборки V, причем $c\to 0$ при $n\to\infty$.

Непараметрическая оценка многомерной плотности вероятности в форме (1) рассмотрена В.А. Епанечниковым [2]. Разработан ряд ядерных оценок плотности вероятности в условиях малых [3 – 6], больших [7 – 12] объемов исходных статистических данных, включая наличие пропусков данных [13, 14] и информации о независимости случайных величин [15 – 17]. На этой основе предложена методика проверки гипотез о тождественности законов распределения случайных величин [18 – 22].

В работе рассматривается методика оптимизации непараметрической оценки плотности вероятности по коэффициентам размытости ядерных функций и использование ее результатов при дискретизации интервала значений случайной величины.

Данное исследование имеет важное значение при формировании критерия Пирсона в задачах проверки гипотез о распределениях случайных величин [23] и построении доверительных интервалов для плотности вероятности на основе ее регрессионной оценки [12].

Оптимизация непараметрической оценки плотности вероятности

Примем критерий оптимальности статистики (1) в виде

$$W_2(c) = \int (\overline{p}(x) - p(x))^2 dx . \qquad (3)$$

Необходимо на основании исходных статистических данных определить наилучшее значение $\overline{c}$ в смысле минимума критерия (3).

Преобразуем с учетом (1) выражение (3)

$$W_2(c) = \frac{1}{n^2 c^2} \sum_{j=1}^{n} \sum_{i=1}^{n} \int \Phi\left(\frac{x - x^j}{c} \right) \Phi\left(\frac{x - x^i}{c} \right) dx -$$

$$- \frac{2}{nc} \sum_{i=1}^{n} \int \Phi\left(\frac{x - x^i}{c} \right) p(x)\, dx + \int p^2(x)\, dx .$$

Заметим, что третий член последнего выражения не зависит от c, поэтому при минимизации $W_2(c)$ его можно не учитывать. Вид второго слагаемого $b(c)$ допускает оценивание статистикой

$$\overline{b}(c) = -\frac{2}{n^2 c} \sum_{j=1}^{n} \sum_{i=1}^{n} \Phi\left(\frac{x^j - x^i}{c} \right). \qquad (4)$$

Причем математические ожидания $b(c)$ и $\overline{b}(c)$ равны, т.е.

$$M(b(c)) = M(\overline{b}(c)),$$

если при формировании (4) выполняется условие $i \neq j$.

Используя методику аналитического исследования непараметрических статистик, предложенную в работе [2] и развитую в статьях [24 – 32], определим

$$M(b(c)) = -\frac{2}{nc} \sum_{i=1}^{n} \int\int \Phi\left(\frac{x - x^i}{c} \right) p(x^i)\, dx^i\, p(x) dx = -\frac{2}{c} \int\int \Phi\left(\frac{x - t}{c} \right) p(t)\, dt\; p(x) dx .$$

При выполнении данных преобразований учитывается, что элементы стати-

стической выборки V являются значениями одной и той же случайной величины t с плотностью вероятности $p(t)$.

Проведем замену переменных $t = x - cu$ и, разлагая функцию $p(x - cu)$ в ряд Тейлора в точке x с учетом свойств ядерной функции (2), при достаточно больших объемах n статистических данных получим следующее асимптотическое выражение: $M(b(c)) \sim -2\int p^2(x)dx - c^2 \int p(x) p^{(2)}(x)dx$, где $p^{(2)}(x)$ – вторая производная плотности вероятности $p(x)$ по x.

Следуя данной технологии преобразований, вычислим

$$M(\overline{b}(c)) = -\frac{2}{n^2 c}\left[\sum_{j=1}^{n}\sum_{\substack{i=1 \\ i\neq j}}^{n}\int\int\Phi\left(\frac{x^j - x^i}{c}\right)p(x^i)dx^i\, p(x^j)dx^j + \right.$$

$$\left. + \sum_{i=1}^{n}\int\int\Phi(0)\, p(x^i)dx^i\, p(x^j)dx^j\right] = \qquad (5)$$

$$= -\frac{2(n-1)}{nc}\int\int\Phi\left(\frac{t - t_1}{c}\right)p(t_1)\,dt_1\; p(t)dt - \frac{2\Phi(0)}{nc}.$$

Нетрудно показать, что первая составляющая выражения (5) при $n \to \infty$ стремится к $M(b(c))$. Поэтому для устранения смещения $M(\overline{b}(c))$ необходимо при вычислении $\overline{b}(c)$ соблюдать условие $i \neq j$. В этом случае в выражении (5) будет отсутствовать вторая составляющая.

Тогда оптимальное значение $\overline{c}$ найдем путем минимизации критерия

$$\overline{W}_2(c) = \frac{1}{n^2 c^2}\sum_{j=1}^{n}\sum_{i=1}^{n}\int\Phi\left(\frac{x - x^j}{c}\right)\Phi\left(\frac{x - x^i}{c}\right)dx - \frac{2}{n^2 c}\sum_{j=1}^{n}\sum_{\substack{i=1 \\ i\neq j}}^{n}\Phi\left(\frac{x^j - x^i}{c}\right). \qquad (6)$$

Методика дискретизации интервала значений случайной величины и ее анализ

При построении непараметрической оценки плотности вероятности (1) используется известное определение $p(x) = dF(x)/dx$ ($F(x)$ – функция распределения случайной величины x) и его разностный аналог с некоторым малым параметром c

$$p(x) \approx (F(x + c) - F(x - c))/(2c) = \frac{1}{2c}\int_{x-c}^{x+c} dF(u) = \frac{1}{c}\int\Phi\left(\frac{x - u}{c}\right)p(u)\,du, \qquad (7)$$

где ядерная функция определяется, например, выражением

$$\Phi\left(\frac{x - u}{c}\right) = \begin{cases} \frac{1}{2}, \text{ если } |x - u| < c, \\ 0, \text{ если } |x - u| \geq c. \end{cases}$$

В результате оценивания математического ожидания ядерной функции (7) по выборке $V = (x^i, i = \overline{1, n})$ получим статистику (1).

Особенность синтеза непараметрической оценки плотности вероятности (1) дает основание связать процедуру дискретизации области изменения случайной

величины x с коэффициентом размытости ядерной функции.

Предлагается количество интервалов дискретизации области значений случайной величины определять как

$$N = \Delta/(2\bar{c}), \tag{8}$$

где $\Delta = \bar{\bar{x}} - \bar{x}$, где $\bar{x} = \min\limits_{i=\overline{1,n}} x^i$, $\bar{\bar{x}} = \max\limits_{i=\overline{1,n}} x^i$.

Проведем сравнение предложенного и ряда традиционных методов дискретизации интервала изменения случайных величин на основе результатов аналитических исследований.

Пусть восстанавливаемая плотность вероятности имеет вид

$$p(x) = \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-x^2}{2}\right).$$

В качестве ядерной функции при оценивании $p(x)$ с помощью статистики (1) будем использовать оптимальное ядро Епанечникова [2]

$$\Phi(u) = \begin{cases} \dfrac{3}{4\sqrt{5}} - \dfrac{3u^2}{20\sqrt{5}} & \forall \quad |u| < \sqrt{5}, \\ 0 & \forall \quad |u| \geq \sqrt{5}. \end{cases}$$

Для выбора количества интервалов дискретизации области изменения значений случайных величин используются предложенный метод (8) и формулы:

Хайнкольда и Гаеде

$$N = \sqrt{n}; \tag{9}$$

Брукса и Каррузера

$$N = 5 \lg n; \tag{10}$$

Старджесса

$$\mathrm{N} = \log_2 n + 1. \tag{11}$$

Известно, что оптимальное значение коэффициента размытости для статистики (1), минимизирующее асимптотическое выражение среднеквадратической ошибки аппроксимации [2], $\frac{1}{nc}\int \Phi^2(u)\,du + \frac{c^4}{4}\int \left(p^{(2)}(x)\right)^2 dx$, определяется формулой

$$c^* = \left[\frac{\int \Phi^2(u)\,du}{n\int \left(p^{(2)}(x)\right)^2 dx}\right]^{\frac{1}{5}}.$$

В соответствии с предложенной методикой интервал Δ значений случайной величины разбивается на равные интервалы дискретизации длиной $2c^*$. С другой стороны, длина интервалов дискретизации при использовании рекомендаций (9), (10), (11) определяется формулой $\beta = \Delta/N$.

Исследуем зависимость $\lambda = 2c^*/\beta$ от объема n исходных статистических

данных для различных методов дискретизации. При этом для принятой ядерной функции и восстанавливаемой плотности вероятности $\int \Phi^2(u)\,du = \frac{3}{5\sqrt{5}}$, $\int \left(p^{(2)}(u)\right)^2\,du = \frac{3}{8\sqrt{\pi}}$. Тогда при использовании формул (9), (10), (11) отношения $\lambda_j,\ \ j=\overline{1,3}$ определяются соответственно выражениями $\lambda_1 = \alpha\sqrt{n}\big/n^{1/5}$, $\lambda_2 = 5\alpha\lg n\big/n^{1/5}$, $\lambda_3 = \alpha\left(\log_2 n+1\right)\big/n^{1/5}$, где $\alpha = \frac{1}{3}\left(\frac{8\sqrt{\pi}}{5\sqrt{5}}\right)^{1/5} \approx 0{,}35$.

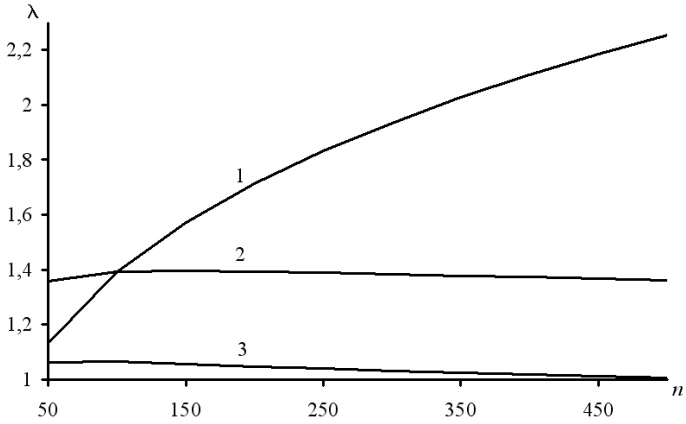

Рис. 1. Зависимость отношений $\lambda_j, j = 1,2,3$ от объема n исходных статистических данных.

Анализ отношений $\lambda_j,\ \ j=\overline{1,3}$ показывает, что длины интервалов дискретизации, полученные с помощью предлагаемого метода и метода Старджесса, сопоставимы (рис. 1). Причем количество интервалов дискретизации, определяемое на их основе, меньше по сравнению с результатами использования формул (9), (10). Отмеченная закономерность особенно характерна для больших объемов n исходных статистических данных.

Предложенный метод допускает его обобщение на решение задачи дискретизации области изменения значений многомерной случайной величины $x = (x_1, x_2, \ldots, x_k)$. Для этого необходимо определить оптимальные коэффициенты размытости $\overline{c}_v$, $v=\overline{1,k}$ непараметрической оценки многомерной оценки плотности вероятности $\overline{p}(x_1,\ldots,x_k) = \left(n\prod_{v=1}^{k} c_v\right)^{-1} \sum_{i=1}^{n}\prod_{v=1}^{k} \Phi\left(\frac{x_v - x_v^i}{c_v}\right)$ из условия минимума статистической оценки среднеквадратического критерия типа (6).

Используя значения $\overline{c}_v$, $v=\overline{1,k}$, можно определить количество интервалов дискретизации компонент x_v $N_v = \Delta_v/(2\overline{c}_v),\ \ v=\overline{1,k}$, где $\Delta_v = \overline{\overline{x}}_v - \overline{x}_v$, $\overline{x}_v = \min\limits_{i=\overline{1,n}} x_v^i$, $\overline{\overline{x}}_v = \max\limits_{i=\overline{1,n}} x_v^i$, $v=\overline{1,k}$.

Заключение

Результаты оптимизации непараметрической оценки плотности вероятности типа Розенблатта – Парзена по коэффициентам размытости ядерных функций являются основой нового метода декомпозиции области изменения значений случайных величин. Его результаты сопоставимы с методом Старджесса и отличаются по сравнению с методами Хайнкольда – Гаеде и Брукса – Каррузера. В отличие

от последних предлагаемый метод основывается не только на использовании объема статистических данных, но и учитывает сведения о восстанавливаемой плотности вероятности, содержащиеся в исходной статистической информации. Перспективность предложенного метода состоит в возможности его обобщения на задачу декомпозиции области изменения случайных величин на многомерные интервалы.

ЛИТЕРАТУРА

1. *Parzen E.* On estimation of a probability density function and mode // Ann. Math. Statistic. – 1962. – Vol. 33, №3. – P.1065-1076.
2. *Епанечников В.А.* Непараметрическая оценка многомерной плотности вероятности // Теория вероятности и ее применения. – 1969. – Т.14, №1. – С.156-161.
3. *Лапко А.В., Лапко В.А., Шарков М.А.* Непараметрические методы обнаружения закономерностей в условиях малых выборок // Изв. вузов. Приборостроение. – 2008. – Т.51, № 8. – С.62-67.
4. *Лапко А.В., Лапко В.А.* Непараметрическая оценка смеси плотностей вероятности, основанная на технологии размножения статистических данных // Вестник СибГАУ. – 2009. – Т.24, № 3. – С.4-6.
5. *Лапко А.В., Лапко В.А.* Непараметрические алгоритмы распознавания образов при случайных значениях коэффициентов размытости ядерных функций // Автометрия. – 2007. – Т.43, № 5. – С.47-55.
6. *Lapko A.V., Lapko V.A.* Nonparametric pattern recognition algorithms for random values of fuzziness factors of kernel functions // Optoelectronics, Instrumentation and Data Processing. – 2007. – Vol. 43, №5. – P.425-432 (DOI: 10.3103/S8756699007050056).
7. *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. – Т.35, № 1. – С.60-64.
8. *Лапко А.В., Лапко В.А.* Свойства непараметрической оценки плотности вероятности многомерных случайных величин в условиях больших выборок // Информатика и системы управления. – 2012. – Т. 32, №2. – С.121-126.
9. *Лапко А.В., Лапко В.А.* Синтез структуры смеси непараметрических оценок плотности вероятности многомерной случайной величины // Системы управления и информационные технологии. – 2011. – Т. 43, №1. – С.12-15.
10. *Лапко А.В., Лапко В.А.* Непараметрические методики анализа множеств случайных величин // Автометрия. – 2003. – Т.39, №1. – С.54-61.
11. *Lapko A.V., Lapko V.A.* Non-parametric Analysis Techniques of Random Values sets // Optoelectronics, Instrumentation and Data Processing. – 2003. – Т.39, №1. – P.44-50.
12. *Лапко А.В., Лапко В.А.* Регрессионная оценка плотности вероятности и ее свойства // Системы управления и информационные технологии. – 2012. – Т.49, №3.1. – С.152-156.
13. *Лапко А.В., Лапко В.А.* Анализ непараметрических алгоритмов распознавания образов в условиях пропуска данных // Автометрия. – 2008. – Т. 44, № 3. – С.65-74.
14. *Lapko A.V., Lapko V.A.* Analysis of nonparametric pattern recognition algorithms under incomplete data // Optoelectronics, Instrumentation and Data Processing. – 2008. – Vol. 44, №3. – P.65-74 (DOI: 10.3103/S8756699008030072).
15. *Лапко А.В., Лапко В.А.* Непараметрическая оценка плотности вероятности независимых случайных величин // Информатика и системы управления. – 2011. – Т.29, №3. – С.118-124.
16. *Лапко А.В., Лапко В.А.* Влияние априорной информации о независимости многомерных случайных величин на свойства их непараметрической оценки плотности вероятности // Системы управления и информационные технологии. – 2012. – Т.48, №2.1. – С.164-167.
17. *Лапко А.В., Лапко В.А.* Свойства непараметрической оценки многомерной плотности веро-

ятности независимых случайных величин // Информатика и системы управления. – 2012. – Т.31, № 1. – С.166-174.
18. *Лапко А.В., Лапко В.А.* Применение непараметрического алгоритма распознавания образов в задаче проверки гипотезы о распределениях случайных величин // Системы управления и информационные технологии. – 2010. – Т. 41, № 3. – С.8-11.
19. *Лапко А.В., Лапко В.А.* Непараметрические алгоритмы распознавания образов в задаче проверки статистической гипотезы о тождественности двух законов распределения случайных величин // Автометрия. – 2010. – Т. 46, № 6. – С.47-53.
20. *Lapko A.V., Lapko V.A.* Nonparametric algorithms of pattern recognition in the problem of testing a statistical hypothesis on identity of two distribution laws of random variables // Optoelectronics, Instrumentation and Data Processing. – 2010. – Vol. 46, №6. – P.545-550 (DOI: 10.3103/S8756699011060069).
21. *Лапко А.В., Лапко В.А.* Сравнение эмпирической и теоретической функций распределения случайной величины на основе непараметрического классификатора // Автометрия. – 2012. – Т.48, № 1. – С.45-49.
22. *Lapko A.V., Lapko V.A.* Comparison of empirical and theoretical distribution functions of a random variable on the basis of a nonparametric classifier // Optoelectronics, Instrumentation and Data Processing. – 2012. – Vol. 48, №1. – P.37-41 (DOI: 10.3103/S8756699012010050).
23. *Лапко А.В., Лапко В.А.* Сравнение непараметрических критериев проверки гипотез о распределениях случайных величин // Вестник СибГАУ. – 2011. – Т.37, № 4. – С.48-52.
24. *Лапко А.В., Лапко В.А.* Анализ асимптотических свойств непараметрической оценки уравнения разделяющей поверхности в двуальтернативной задаче распознавания образов // Автометрия. – 2010. – Т.46, № 3. – С.48-53.
25. *Lapko A.V., Lapko V.A.* Analysis of Asymptotic Properties of Nonparametric Estimate of the Equation of the Separation Surface in a Two-Alternative Problem of Pattern Recognition // Optoelectronics, Instrumentation and Data Processing. – 2010. – Vol. 46, №3. – P.243-247.
26. *Лапко А.В., Лапко В.А.* Разработка и исследование двухуровневых непараметрических систем классификации // Автометрия. – 2010. – Т.46, № 1. – С.70-78.
27. Lapko A.V., Lapko V.A. Development and Investigation of Two-Level Non-Parametric Estimators // Optoelectronics, Instrumentation and Data Processing. – 2010. – Vol. 46, №1. – P.56-63 (DOI: 10.3103/S8756699010010073).
28. *Лапко А.В., Лапко В.А.* Синтез структуры семейства непараметрических решающих функций в задаче распознавания образов // Автометрия. – 2011. – Т. 47, № 4. – С.76-82.
29. *Lapko A.V., Lapko V.A.* Synthesis of the Structure of a Family of Nonparametric Decision Functions in the Pattern Recognition Problem // Optoelectronics, Instrumentation and Data Processing. – 2011. – Vol. 47, №4. – P.383-387 (DOI: 10.3103/S8756699011040091).
30. *Лапко А.В., Лапко В.А.* Коллектив непараметрических решающих функций в двуальтернативной задаче распознавания образов // Системы управления и информационные технологии. – 2009. – Т.37, № 3.1. – С.156 – 160.
31. *Лапко А.В., Лапко В.А.* Асимптотические свойства многомерной непараметрической оценки уравнения разделяющей поверхности в двуальтернативной задаче распознавания образов // Системы управления и информационные технологии. – 2010. – Т.39, №1. – С.16-19.
32. *Лапко А.В., Лапко В.А.* Непараметрическая оценка уравнения разделяющей поверхности в условиях больших выборок и ее свойства // Системы управления и информационные технологии. – 2010. – Т.39, № 1.2. – С. 300-304.

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