УДК 519.7

МЕТОД ДЕКОМПОЗИЦИИ ИНТЕРВАЛА ЗНАЧЕНИЙ СЛУЧАЙНЫХ ВЕЛИЧИН, ОСНОВАННЫЙ НА РЕЗУЛЬТАТАХ ОПТИМИЗАЦИИ НЕПАРАМЕТРИЧЕСКОЙ ОЦЕНКИ ПЛОТНОСТИ ВЕРОЯТНОСТИ*

А. В. Лапко[1, 2], **В. А. Лапко**[1, 2]

[1] *Институт вычислительного моделирования СО РАН, 660036, г. Красноярск, Академгородок, 50, стр. 44*
[2] *Сибирский государственный аэрокосмический университет им. академика М. Ф. Решетнёва, 660014, г. Красноярск, просп. им. Газеты «Красноярский рабочий», 31*
E-mail: lapko@icm.krasn.ru

Предлагается новый метод декомпозиции интервала значений случайных величин, основанный на результатах оптимизации непараметрической оценки плотности вероятности типа Розенблатта — Парзена. Рассматривается его применение в задаче проверки гипотезы тождественности законов распределения двух последовательностей одномерных случайных величин.

Ключевые слова: метод декомпозиции, критерий Пирсона, проверка гипотез, непараметрическая оценка, плотность вероятности, вычислительный эксперимент.

Введение. Критерий Пирсона остаётся пока единственным эффективным методом проверки гипотез о распределениях многомерных случайных величин [1]. Его применение позволяет проверять гипотезы о тождественности эмпирического и гипотетического законов распределения, а также гипотезы о совпадении распределений в двух выборках случайных величин. Однако методика формирования критерия Пирсона содержит трудно формализуемый этап разбиения области возможных значений случайной величины на интервалы.

Данный этап отсутствует в критерии Колмогорова — Смирнова, основанном на анализе преобразований разности функций распределения одномерных случайных величин [2].

Для обхода проблемы декомпозиции области изменения значений случайных величин в задачах проверки гипотез об их распределении в работах [3–7] предлагается использовать непараметрические алгоритмы распознавания образов [8–10]. Идея такого подхода состоит в обосновании замены задачи сравнения законов распределения случайных величин проверкой гипотезы о равенстве статистической оценки вероятности ошибки распознавания образов и определённого порогового значения. Перспективность указанного направления заключается в возможности его обобщения на задачи проверки гипотез о распределении многомерных случайных величин.

Цель данной работы состоит в обосновании и исследовании метода декомпозиции интервала значений одномерной случайной величины, основанного на результатах оптимизации непараметрической оценки плотности вероятности ядерного типа.

*Работа выполнена в рамках базовой части государственного задания Министерства образования и науки РФ высшим учебным заведениям на 2014–2016 гг. (СибГАУ № Б121/14).

Непараметрическая оценка плотности вероятности. Пусть $V = (x^i,\ i = \overline{1,n})$ — выборка из n независимых наблюдений одномерной случайной величины с неизвестной плотностью вероятности $p(x)$.

В качестве приближения по эмпирическим данным искомой плотности $p(x)$ примем её непараметрическую оценку [11]

$$\bar{p}(x) = \frac{1}{nc}\sum_{i=1}^{n}\Phi\Big(\frac{x - x^i}{c}\Big), \tag{1}$$

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

$$\begin{gathered}\Phi(u) = \Phi(-u); \quad 0 \le \Phi(u) < \infty; \quad \int\limits_{-\infty}^{+\infty}\Phi(u)du = 1; \\ \int\limits_{-\infty}^{+\infty}u^2\Phi(u)du = 1; \quad \int\limits_{-\infty}^{+\infty}u^m\Phi(u)du < \infty \quad \text{при}\, 0 \le m < \infty.\end{gathered} \tag{2}$$

Коэффициент размытости $c = c(n)$ ядерных функций непараметрической оценки (1) убывает с увеличением объёма n выборки V, причём $c \to 0$ при $n \to \infty$. Его оптимальное значение, минимизирующее асимптотическое выражение среднеквадратической ошибки аппроксимации [12–14]

$$\frac{1}{nc}\int\limits_{-\infty}^{+\infty}\Phi^2(u)du + \frac{c^4}{4}\int\limits_{-\infty}^{+\infty}(p^{(2)}(x))^2dx,$$

определяется формулой

$$c^* = \left[\int\limits_{-\infty}^{+\infty}\Phi^2(u)du\Big/\Big(n\int\limits_{-\infty}^{+\infty}(p^{(2)}(x))^2dx\Big)\right]^{1/5}, \tag{3}$$

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

Формула (3) имеет теоретическое значение и используется при исследовании аппроксимационных свойств статистики (1) [15].

Выбор коэффициента размытости c по исходным данным V будем осуществлять из условия экстремума статистического критерия $L(c)$, определяющего качество оценивания плотности вероятности $p(x)$ её непараметрической оценкой $\bar{p}(x)$. Например, из условия максимума статистической оценки функции правдоподобия [16]

$$L(c) = \prod_{j=1}^{n}\bar{p}(x^j), \tag{4}$$

где

$$\bar{p}(x^j) = \frac{1}{(n-1)c}\sum_{\substack{i=1\\ i\ne j}}^{n}\Phi\Big(\frac{x^j - x^i}{c}\Big).$$

Применение результатов оптимизации $\bar{p}(x)$ при проверке гипотезы о распределении случайных величин. Пусть X_1 и X_2 — две генеральные совокупности с произвольными законами распределения. Необходимо по независимым выборкам $V_1 = (x^i,\ i = \overline{1, n_1})$ и $V_2 = (x^i,\ i = \overline{1, n_2})$, извлечённым из данных генеральных совокупностей, проверить гипотезу $H_0\colon P(X_1) \equiv P(X_2)$ о тождественности их законов распределения.

Зададим количество N интервалов дискретизации области значений случайной величины x по исходным данным $V = V_1 \cup V_2 = (x^i,\ i = \overline{1, n})$, $n = n_1 + n_2$. Для этого осуществим синтез непараметрической оценки $\bar{p}(x)$ (1) плотности вероятности и по приведённой выше методике найдём оптимальный коэффициент размытости $\bar{c}$ ядерных функций.

Тогда количество интервалов дискретизации выражается как $N = \Delta/(2\bar{c})$. Здесь $\Delta = \bar{\bar{x}} - \bar{x}$, где $\bar{x} = \min\limits_{i=\overline{1,n}} x^i$, $\bar{\bar{x}} = \max\limits_{i=\overline{1,n}} x^i$.

Вычислим частоты $\bar{P}_1^j$, $\bar{P}_2^j$, $\bar{P}_{12}^j$ попадания элементов последовательностей случайных величин V_1, V_2 и $V_1 \cup V_2$ в каждый j-й интервал ($j = \overline{1, N}$).

Определим значение случайной величины [1, с. 330]

$$Z = \sum_{t=1}^{2} n_t \sum_{j=1}^{N} (\bar{P}_t^j - \bar{P}_{12}^j)^2 / \bar{P}_{12}^j,$$

которое имеет χ^2-распределение с $k = N - 1$ степенями свободы.

По таблице χ^2-распределения найдём порог $\chi^2(k, \alpha)$ критерия Пирсона при значении k и заданном уровне значимости α. Гипотеза H_0 справедлива, если $Z < \chi^2(k, \alpha)$, иначе она отвергается.

Анализ результатов вычислительных экспериментов. Проведём сравнение эффективности предложенного и ряда общепризнанных методов дискретизации интервала значений случайных величин при решении задачи проверки гипотезы о тождественности их законов распределения. В качестве методики проверки гипотезы примем критерий Пирсона. Последовательности случайных величин $V_1 = (x^i,\ i = \overline{1, n_1})$ и $V_2 = (x^i,\ i = \overline{1, n_2})$ формируются на основе датчиков с нормальным $N(1{,}5; 0{,}45)$ и равномерным ($x \in [0; 3]$) законами распределения.

Для выбора количества интервалов дискретизации области изменения значений случайных величин используются предложенный метод и формулы Старджесса, Брукса и Каррузера, Хайнкольда и Гаеде соответственно:

$$N = \log_2 n + 1, \tag{5}$$

$$N = 5 \lg n, \tag{6}$$

$$N = \sqrt{n}, \tag{7}$$

где $n = n_1 + n_2$.

Синтез непараметрической оценки плотности вероятности (1) осуществляется на основе ступенчатой ядерной функции

$$\Phi(u) = \begin{cases} 1/2 & \forall |u| < 1, \\ 0 & \forall |u| \geq 1 \end{cases} \tag{8}$$

и оптимального ядра Епанечникова [12]

$$\Phi(u) = \begin{cases} \dfrac{3}{4\sqrt{5}} - \dfrac{3u^2}{20\sqrt{5}} & \forall |u| < \sqrt{5}, \\ 0 & \forall |u| \geq \sqrt{5}. \end{cases} \tag{9}$$

Вычислительные эксперименты при фиксированных условиях исследования повторяются 100 раз. По полученным результатам оценивается вероятность $\bar{P}_0$ выполнения гипотезы H_0 о тождественности законов распределения случайных величин на основе исследуемых методов дискретизации. Риск α отвергнуть гипотезу H_0 принимается равным 0,05.

Зависимости оценок $\bar{P}_0$ вероятностей справедливости гипотезы H_0 от объёма n исходных данных при $n_1 = n_2$ в условиях сравнения двух априори тождественных законов и в условиях нормального и равномерного законов распределения случайных величин отражены в табл. 1 и 2. Результаты, представленные в столбцах С, Б, Г, О, получены при использовании соответственно формул (5)–(7) и предложенного метода дискретизации. В столбцах О(8), О(9) приведены оценки $\bar{P}_0$ вероятностей выполнения гипотезы H_0 в условиях применения ядерных функций (8), (9). Анализ результатов вычислительных экспериментов показывает, что эффективность критерия Пирсона сопоставима при использовании исследуемых методов декомпозиции интервала значений случайных величин. Отмеченная закономерность свойственна условиям сравнения априори тождественных и разных законов распределения (см. табл. 1 и 2).

При сравнении последовательностей случайных величин с равномерным и нормальным законами распределения в условиях $n > 150$ применение исследуемых методов дискретизации обеспечивает безошибочное отклонение гипотезы H_0. При $n < 50$ результаты их использования неудовлетворительные. Этот факт объясняется неравномерностью распределения исходных данных в области изменения случайных величин, что приводит к снижению качества оценивания вероятности принадлежности значений случайных величин интервалам дискретизации. В таких условиях максимум функции правдоподобия смещается в область бóльших значений коэффициентов размытости ядерных функций, что приводит к снижению количества интервалов дискретизации и эффективности исследуемых методов.

Таблица 1

n	Равномерные законы распределения					n	Нормальные законы распределения				
	С	Б	Г	О(8)	О(9)		С	Б	Г	О(8)	О(9)
10	1	1	1	1	1	10	0,06	0,05	0,16	0,06	0,95
30	0,97	0,96	1	0,96	0,98	30	0,38	0,28	0,43	0,33	0,98
50	0,93	0,96	0,93	0,92	0,89	50	0,63	0,58	0,63	0,57	0,93
70	0,93	0,95	0,95	0,96	0,96	70	0,91	0,87	0,88	0,84	0,97
90	0,94	0,98	0,97	0,98	0,95	90	0,98	0,92	0,90	0,86	0,95
110	0,92	0,90	0,90	0,91	0,99	110	0,99	0,97	0,97	0,98	0,95
130	0,95	0,94	0,94	0,92	0,97	130	1	1	1	0,97	0,94
150	0,92	0,92	0,94	0,93	0,92	150	1	1	1	1	0,96
170	0,96	0,92	0,96	0,96	0,96	170	1	1	1	0,97	0,99
190	0,99	0,97	0,98	0,95	0,98	190	1	1	1	1	0,92
210	0,96	0,96	0,96	0,95	0,96	210	1	1	1	1	0,97
230	0,97	0,96	0,95	0,96	0,94	230	1	1	1	1	0,98
250	0,99	0,96	0,96	0,96	0,96	250	1	1	1	1	0,97
270	0,97	0,97	0,97	0,97	0,98	270	1	1	1	1	0,99
290	0,95	0,94	0,96	0,95	0,97	290	1	1	1	1	0,93

Таблица 2

n	С	Б	Г	О(8)	О(9)
10	0,94	0,96	0,89	0,93	0,95
30	0,58	0,66	0,55	0,7	0,6
50	0,49	0,39	0,49	0,53	0,4
70	0,12	0,16	0,12	0,15	0,15
90	0,02	0,07	0,05	0,08	0,11
110	0,02	0,03	0,03	0,04	0,04
130	0,01	0,01	0,01	0,01	0,03
150	0	0	0	0	0
170	0	0	0	0	0
190	0	0	0	0	0
210	0	0	0	0	0
230	0	0	0	0	0
250	0	0	0	0	0
270	0	0	0	0	0
290	0	0	0	0	0

Выбор вида ядерной функции оказывает несущественное влияние на эффективность критерия Пирсона.

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 $\bar{c}_v$, $v = \overline{1,k}$, непараметрической оценки многомерной плотности вероятности

$$\bar{p}(x_1, \ldots, x_k) = \Big(n \prod_{v=1}^{k} \bar{c}_v\Big)^{-1} \sum_{i=1}^{n} \prod_{v=1}^{k} \Phi\Big(\frac{x_v - x_v^i}{\bar{c}_v}\Big)$$

из условия максимума критерия $L(c_1, \ldots, c_k) = \prod\limits_{j=1}^{n} \bar{p}(x_1^j, \ldots, x_k^j)$.

Составляющие $L(c_1, \ldots, c_k)$ вычисляются в соответствии с выражением

$$\bar{p}(x_1^j, \ldots, x_k^j) = \Big((n-1) \prod_{v=1}^{k} \bar{c}_v\Big)^{-1} \sum_{\substack{i=1 \\ i \neq j}}^{n} \prod_{v=1}^{k} \Phi\Big(\frac{x_v^j - x_v^i}{\bar{c}_v}\Big).$$

Используя значения $\bar{c}_v$, $v = \overline{1,k}$, определим количество интервалов дискретизации компонент x_v:

$$N_v = \Delta_v/(2\bar{c}_v), \quad v = \overline{1,k},$$

где $\Delta_v = \bar{\bar{x}}_v - \bar{x}_v$ $\Big(\bar{x}_v = \min\limits_{i=\overline{1,n}} x_v^i$, $\bar{\bar{x}}_v = \max\limits_{i=\overline{1,n}} x_v^i$, $v = \overline{1,k}\Big)$.

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 применение в критерии Пирсона при проверке гипотез о распределении одномерных случайных величин позволяет получить показатели эффективности, сопоставимые с известными методами декомпозиции. В отличие от последних предлагаемый метод основывается не только на объёме статистических данных, но и учитывает сведения о сравниваемых законах распределения случайных величин, содержащихся в исходной статистической информации.

Перспективность предложенного метода состоит в возможности его обобщения на задачу декомпозиции области изменения значений случайных величин на многомерные интервалы.

СПИСОК ЛИТЕРАТУРЫ

1. **Пугачев В. С.** Теория вероятностей и математическая статистика. М.: Наука, 1979. 496 с.
2. **Смирнов Н. В.** Оценка расхождения между кривыми распределения в двух независимых выборках // Бюлл. Моск. университета. 1930. **2**, № 2. С. 3–14.
3. **Лапко А. В., Лапко В. А.** Непараметрические алгоритмы распознавания образов в задаче проверки статистической гипотезы о тождественности двух законов распределения случайных величин // Автометрия. 2010. **46**, № 6. С. 47–53.
4. **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. **41**, № 3. С. 8–11.
5. **Лапко А. В., Лапко В. А.** Непараметрические алгоритмы распознавания образов в задаче проверки гипотезы о распределениях случайных величин // Изв. вузов. Приборостроение. 2011. **54**, № 4. С. 67–72.
6. **Лапко А. В., Лапко В. А., Струков И. И., Гусаров А. А.** Непараметрический классификатор и критерий Колмогорова в задаче сравнения эмпирической и теоретической функций распределения одномерной случайной величины // Вестн. СибГАУ. 2011. **35**, № 2. С. 37–40.
7. **Лапко А. В., Лапко В. А.** Сравнение эмпирической и предлагаемой функций распределения случайной величины на основе непараметрического классификатора // Автометрия. 2012. **48**, № 1. С. 45–49.
8. **Лапко А. В., Лапко В. А.** Разработка и исследование двухуровневых непараметрических систем классификации // Автометрия. 2010. **46**, № 1. С. 70–78.
9. **Лапко А. В., Лапко В. А.** Синтез структуры семейства непараметрических решающих функций в задаче распознавания образов // Автометрия. 2011. **47**, № 4. С. 76–82.
10. **Лапко А. В., Лапко В. А.** Анализ асимптотических свойств непараметрической оценки уравнения разделяющей поверхности в двухальтернативной задаче распознавания образов // Автометрия. 2010. **46**, № 3. С. 48–53.
11. **Parzen E.** On estimation of a probability density function and mode // Ann. Math. Stat. 1962. **33**, N 3. P. 1065–1076.
12. **Епанечников В. А.** Непараметрическая оценка многомерной плотности вероятности // Теория вероятностей и ее применения. 1969. **14**, № 1. С. 156–161.
13. **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. **35**, № 1. С. 60–64.

14. **Лапко А. В., Лапко В. А.** Свойства непараметрической оценки плотности вероятности многомерных случайных величин в условиях больших выборок // Информатика и системы управления. 2012. **32**, № 2. С. 121–126.
15. **Лапко А. В., Лапко В. А.** Анализ дисперсии среднеквадратической ошибки аппроксимации непараметрической оценки плотности вероятности ядерного типа // Информатика и системы управления. 2012. **33**, № 3. С. 132–139.
16. **Деврой Л., Дьерфи Л.** Непараметрическое оценивание плотности. L_1-подход: Пер. с англ. М.: Мир, 1988. 408 с.

Поступила в редакцию 24 мая 2013 г.