

УДК 004.942

МЕТОДИКА ОЦЕНИВАНИЯ ПАРАМЕТРОВ СЛУЧАЙНОЙ ВЕЛИЧИНЫ СО СМЕШАННЫМ ДВУМОДАЛЬНЫМ ЗАКОНОМ РАСПРЕДЕЛЕНИЯ НА ОСНОВЕ СОВМЕСТНОГО ИСПОЛЬЗОВАНИЯ АППРОКСИМАЦИИ РОЗЕНБЛАТТА – ПАРЗЕНА, МЕТОДА МНИМЫХ ИСТОЧНИКОВ И ГЕНЕТИЧЕСКИХ АЛГОРИТМОВ

Поршнев С.В., Копосов А.С.

ФГАОУ ВПО «Уральский федеральный университет имени первого Президента России Б.Н. Ельцина», Институт радиоэлектроники и информационных технологий, Екатеринбург, e-mail: alexkopas@gmail.com

В статье обсуждаются результаты совместного использования аппроксимации Розенблатта – Парзена и метода мнимых источников для оценки параметров случайной величины со смешанным двумодальным распределением. В исследовании изучались случайные последовательности, генерируемые в соответствии с задаваемыми двумодальными законами распределения. На первом этапе вычислялись оценки функции плотности распределения с помощью аппроксимации Розенблатта – Парзена, по которым оценивались математические ожидания. На втором этапе с помощью генетических алгоритмов, в которых целевая функция задавалась в соответствии с методом мнимых источников, вычислялись оценки остальных параметров. Точность нахождения параметров двумодальных распределений оценивалась на основе сравнения теоретических параметров распределений и соответствующих параметров, полученных экспериментально. Анализ результатов применения предложенной методики, основанной на совместном использовании аппроксимации Розенблатта – Парзена, метода мнимых источников и генетических алгоритмов, для оценки значений параметров двумодальных распределений позволяет сделать следующие выводы: при использовании генетических алгоритмов для нахождения одновременно всех девяти параметров изучаемого распределения погрешности оценок превышают 50%, что является следствием высокой размерности решаемой задачи. Предложена методика последовательного нахождения параметров двумодального распределения.

Ключевые слова: функция распределения, аппроксимация Розенблатта – Парзена, генетический алгоритм, двумодальное распределение, метод мнимых источников

TECHNIQUE OF ESTIMATION OF PARAMETERS OF RANDOM VALUE WITH MIXED BIMODAL DISTRIBUTION LAW BASED ON COMBINED USAGE OF ROZENBLATT-PARZEN APPROXIMATION, IMAGINARY SOURCES METHOD AND GENETIC ALGORITHMS

Porshnev S.V., Koposov A.S.

Ural Federal University named after the first President of Russia B.N. Yeltsin, Institute of Radioelectronics and Information Technologies, Ekaterinburg, e-mail: alexkopas@gmail.com

In this article are considered the results of combined usage of Rozenblatt – Parzen approximation and imaginary sources method for estimation of the parameters of random value with mixed bimodal distribution law. We have examined random sequences with mixed distribution law containing two peaks. At the first stage we calculated Rozenblatt – Parzen approximation, which gives estimation of mean values. At the second stage using genetic algorithms with fitness function defined in accordance with imaginary sources method we solved for other parameters. Estimation of calculation accuracy of parameters was performed by comparison of theoretical parameters with experimental ones. The analysis of proposed method application results allows us to make the following conclusions: using genetic algorithms for solving all nine parameters simultaneously produces a 50% error. This is the result from a high-dimensional problem. We have proposed a technique of a sequential solving for distribution parameters.

Keywords: cumulative distribution function, Rozenblatt – Parzen approximation, genetic algorithms, bimodal distribution, imaginary sources method

Восстановление функции распределения по выборке случайных данных, полученных в результате проведения тех или иных экспериментов, является основной задачей математической статистики [1], которая имеет важное практическое значение, например, при решении задач прочностной надежности элементов и объектов нефтегазового оборудования [2]. Обсуждаемая задача имеет следующую постановку: по экспериментальной выборке значений случайной величины $X_i, i = 1, N$ из генеральной со-

вокупности найти функцию распределения $F(y) = \Pr\{X \leq y\}$, связанную с плотностью распределения $f(y)$ интегральным отношением:

$$F(y) = \int_{-\infty}^y f(\xi) d\xi. \quad (1)$$

Известны два основных подхода к решению этой задачи: параметрический и непараметрический.

Параметрический подход предусматривает выбор на основе той априорной

информации вида функции распределения случайной величины $F(y)$, зависящей от некоторого набора параметров, и получении оценок их значений по имеющейся выборке данных, обеспечивающих максимальную близость теоретической функции распределения $F(y)$ и эмпирической функции распределения

$$F_N(y) = \frac{1}{N} \sum_{i=1}^N \Theta(y - x_i), \quad (2)$$

где функция Хэвисайда

$$\Theta(y - x_i) = \begin{cases} 1, & \text{при } y - x_i \geq 0, \\ 0, & \text{при } y - x_i < 0 \end{cases}$$

в соответствии с выбранной мерой близости, зависящей, вообще говоря, от вида распределения [3].

Существование решения обсуждаемой задачи обеспечивает центральная теорема математической статистики, согласно которой с ростом объема выборки N функция $F_N(y)$ с вероятностью, равной единице, равномерно приближается к $F(y)$:

$$\Pr \left\{ \limsup_{N \rightarrow \infty} |F_N(y) - F(y)| = 0 \right\} = 1.$$

В основе непараметрической статистики лежит подход, позволяющий получать адаптивные оценки эмпирических распределений в виде некоторых функционалов, не зависящих от вида неизвестного априорного распределения [4]. Для восстановления неизвестной функции распределения в непараметрической статистике известен ряд методов и алгоритмов [4]: метод гистограмм, «гребенка», метод ближайших

соседей, метод разложения по базисным функциям, аппроксимация Розенблатта – Парзена и ряд других. Работоспособность методов непараметрической статистики и целесообразность их применения при анализе экспериментальных данных подтверждается результатами, полученными различными исследователями (см. список литературы к разделу «Введение» в [4]). Например, в [2] показано, что аппроксимация Розенблатта – Парзена оказывается весьма эффективной в задаче оценки долговечности нефте- и газопроводов на основе анализа накопленной статистической информации.

Результаты исследования особенностей аппроксимации Розенблатта – Парзена в задаче аппроксимации одномодальных распределений дискретных и непрерывных случайных величин с ограниченной областью изложены в [5, 6] соответственно.

В связи с тем, что на практике, например при оценке прочностной надежности изделий [2] или анализе суточной выработки экскаваторов на горных работах [3], требуется получение оценок распределений случайных величин с двумодальными законами распределения, разработка методов оценки их параметров является актуальной задачей.

В данной статье приведено описание и обоснование методики оценки параметров двумодального распределения случайной величины, каждая мода которого имеет нормальный закон распределения с ограниченной областью рассеяния.

Функция изучаемого распределения записывается в следующем виде:

$$F(x) = F_1(x, \mu_1, \sigma_1, a_1, b_1) \cdot \alpha + F_2(x, \mu_2, \sigma_2, a_2, b_2) \cdot (1 - \alpha), \quad (3)$$

где μ_1 – математическое ожидание первой составляющей; σ_1 – математическое ожидание первой составляющей; a_1, b_1 – границы области рассеяния первой составляющей; μ_2 – математическое ожидание первой составляющей; σ_2 – математическое ожидание первой составляющей; a_2, b_2 – границы области рассеяния первой составляющей; α – доля первой составляющей в общем распределении.

Пример двумодальной функции распределения случайной величины, каждая мода которого имеет нормальный закон распределения с ограниченной областью рассеяния, представлен на рис. 1.

Отметим, что в общем случае задача оценки параметров распределения сводится к решению той или иной системы нелинейных уравнений, для которых в подавляющем большинстве случаев приходится

использовать соответствующие численные методы, например, итерационный метод Ньютона. Однако сходимость интегральной последовательности к истинному решению оказывается очень сильно зависящей от выбора начального приближения. Вследствие этого в рассматриваемой задаче представляется перспективным использовать эвристические методы случайного поиска, результативность которых, как утверждается, не зависит от начального приближения и позволяет найти оптимальное решение при любых начальных условиях. Одним из таких методов являются генетические алгоритмы (ГА) [7], которые были использованы в проведенном исследовании.

На первом этапе исследования была изучена возможность одновременного получения оценок всех значений параметров

двумодального распределения $\mu_1, \sigma_1, a_1, b_1, \mu_2, \sigma_2, a_2, b_2, \alpha$ с помощью ГА, в котором в качестве целевой функции использовались среднеквадратические отклонения отсортированной по возрастанию исходной последовательности и отсортированной по возрастанию последовательности, сгенерированной в соответствии с экспериментальным законом распределения (3):

$$fitness = \frac{1}{N} \sum_{i=1}^N (x_i - x_i^*)^2, \quad (4)$$

где N – число элементов в исходной выборке; x – исходная выборка; x^* – выборка, сгенерированная с помощью метода обратного преобразования в соответствии с законом (3) и текущими значениями параметров распределения, задаваемых ГА.

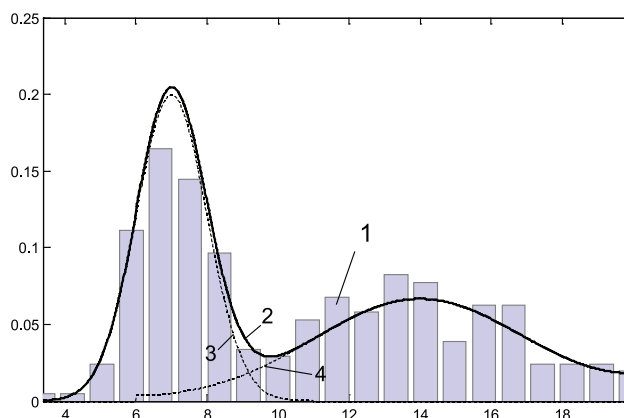


Рис. 1. Пример двумодального распределения:
1 – гистограмма случайной последовательности; 2 – плотность распределения случайной величины; 3 – график функции $dF_1(x, \mu_1, \sigma_1, a_1, b_1)/dx$; 4 – график функции $dF_2(x, \mu_2, \sigma_2, a_2, b_2)/dx$

Методика оценивания параметров двумодального распределения

Результаты многочисленных вычислительных экспериментов, проведенных при различных настройках ГА, показали, что получаемые значения были далеки от истинных – средняя погрешность вычислений составила более 50%. Этот результат, с нашей точки зрения, объясняется высокой размерностью пространства, в котором ищется минимум функции (7). Таким образом, прямое решение данной задачи с помощью ГА не является приемлемым. В этой

связи возникла необходимость разработки методики решения рассматриваемой задачи, позволяющей уменьшить размерность пространства, в котором ищется минимум функции (4), за счет оценки одного или нескольких параметров, не используя ГА.

Для реализации данной идеи было предложено использовать метод Розенблатта – Парзена, позволяющий аппроксимировать функции плотности распределения [8, 9, 6]. Результаты применения метода Розенблатта – Парзена к случайной последовательности с двумодальным распределением представлен на рис. 2.

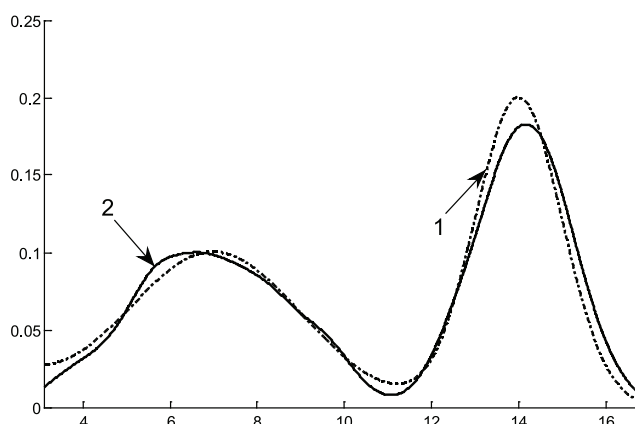


Рис. 2. Плотность двумодального распределения:
1 – теоретическая плотность двумодального распределения; 2 – аппроксимация Розенблатта – Парзена плотности распределения случайной последовательности, сгенерированной в соответствии с теоретическим законом распределения

Из рис. 2 видно, что на основе анализа аппроксимирующей кривой можно получать оценки математических ожиданий каждой из мод исходного распределения $\mu_{1,2}$, а также левую границу первой составляющей a_1 и правую границу второй составляющей b_2 . Таким образом, с помощью ГА остается вычислить следующие параметры распределения $\sigma_1, b_1, \sigma_2, a_2, \alpha$, т.е. удается уменьшить размерность задачи с 9 до 5.

Таким образом, методика нахождения параметров двумодального распределения случайной последовательности $\{x_i\}, i=\overline{1,N}$ (3) реализуется выполнением следующей последовательности действий:

1. Вычисление в соответствии с методом Розенблатта – Парзена значений функции, аппроксимирующей плотность распределения (3), $-F_{RP}(x)$.

2. Вычисление оценок значений параметров $\tilde{\mu}_1, \tilde{\mu}_2$ – абсцисс локальных максимумов функции $F_{RP}(x)$.

3. Вычисление оценки левой границы области рассеяния моды распределения (3), описываемой функцией

$$F_1(x, \mu_1, \sigma_1, a_1, b_1): \tilde{a}_1 = \min_i (X_i).$$

$$numT = numS \times numM \times numC \times numCfr \times numP = 3 \cdot 1 \cdot 5 \cdot 3 \cdot 3 = 135$$

различных комбинаций настроек. Для каждого набора настроек было выполнено 50 независимых испытаний. Для повышения достоверности в качестве значений параметров $\tilde{\sigma}_1, \tilde{b}_1, \tilde{\sigma}_2, \tilde{a}_2, \tilde{\alpha}$ принимались средние по ансамблю испытаний значения.

Для количественной оценки качества найденных значений параметров $\tilde{\sigma}_1, \tilde{b}_1, \tilde{\sigma}_2, \tilde{a}_2$ решения мы использовали величину Δ_x :

$$\Delta_y = \frac{\max |y_{teor} - \tilde{y}|}{L} \cdot 100\%, \quad (5)$$

где $\tilde{y} = \tilde{\sigma}_1, \tilde{b}_1, \tilde{\sigma}_2, \tilde{a}_2$.

Для количественной оценки качества найденного значения параметра $\tilde{\alpha}$ мы использовали величину Δ_α :

$$\Delta_\alpha = \frac{|\alpha_{teor} - \tilde{\alpha}|}{\alpha_{max}} \cdot 100\%, \quad (6)$$

где $\alpha_{max} = 1$.

Результаты расчетов погрешностей параметров $\tilde{\sigma}_1, \tilde{b}_1, \tilde{\sigma}_2, \tilde{a}_2, \tilde{\alpha}$ в соответствии с (5),

$$\Delta_{integr} = \frac{\sum_i^N \left(F_{teor}(x, \mu_1, \sigma_1, a_1, b_1, \mu_2, \sigma_2, a_2, b_2, \alpha) - F_{pract}(x, \tilde{\mu}_1, \tilde{\sigma}_1, \tilde{a}_1, \tilde{b}_1, \tilde{\mu}_2, \tilde{\sigma}_2, \tilde{a}_2, \tilde{b}_2, \tilde{\alpha}) \right)^2}{\sum_i^N F_{teor}(x, \mu_1, \sigma_1, a_1, b_1, \mu_2, \sigma_2, a_2, b_2, \alpha)^2},$$

4. Вычисление оценки правой границы области рассеяния моды распределения (3), описываемой функцией

$$F_2(x, \mu_2, \sigma_2, a_2, b_2): \tilde{b}_2 = \max_i (X_i).$$

5. Вычислить, используя ГА, значения параметров $\sigma_1, b_1, \sigma_2, a_2, \alpha$.

В ходе проведенных исследований мы использовали следующие настройки ГА:

- селекция: $S = \langle \text{равномерная; турнирная; на основе рулетки} \rangle$;

- мутация: $M = \langle \text{адаптивная} \rangle$, т.к. в задаче присутствуют ограничения;

- кроссовер: $C = \langle \text{одноточечный; двухточечный; усредненный; разбросанный, эвристический} \rangle$;

- доля кроссовера: $Cfr = \langle 0,3; 0,6; 0,9 \rangle$ (оставшаяся часть приходилась на мутацию);

- размер популяции: $P = \langle 5, 15, 30 \rangle$.

Таким образом, каждому из проведенных экспериментов можно поставить в соответствие определенный набор настроек – кортеж $\langle S, M, C, Cfr, P \rangle$. Всего в проведенных экспериментах было рассмотрено

(6) для тестовых последовательностей, генерируемых в соответствии с (3), где $\mu_1 = 7, \sigma_1 = 2, a_1 = 3, b_1 = 14, \mu_2 = 714, \sigma_2 = 1, a_2 = 6, b_2 = 18, \alpha = 0,5$, представлены на рис. 3.

Из рис. 3 видно, что виды селекции и кроссовера не оказывают существенного влияния на погрешность вычисления параметров $\tilde{\sigma}_1, \tilde{b}_1, \tilde{\sigma}_2, \tilde{a}_2, \tilde{\alpha}$. В то время как при увеличении доли кроссовера и числа особей в популяции точность вычисления ключевых параметров $\sigma_1, \sigma_2, \alpha$ увеличивается при одновременном увеличении времени расчетов. Приведем средние значения погрешностей параметров:

$$\overline{\Delta_{\sigma_1}} = 3,79\%; \quad \overline{\Delta_{\sigma_2}} = 1,84\%;$$

$$\overline{\Delta_{b_1}} = 15,65\%; \quad \overline{\Delta_{a_2}} = 10,45\%;$$

$$\overline{\Delta_\alpha} = 9,25\%.$$

Для интегрального оценивания качества предложенной методики был использован следующий показатель:

значения которого в проведенных вычислительных экспериментах изменялись в диапазоне

$$\left[(\Delta_{integr})_{\min}, (\Delta_{integr})_{\max} \right],$$

$\langle S = \text{турнирная}, M = \text{адаптивная}, C = \text{эвристический}, Cfr = 0,6, P = 30 \rangle$,

наихудший $(\Delta_{integr} = (\Delta_{integr})_{\max})$ –

$\langle S = \text{рулетка}, M = \text{адаптивная}, C = \text{одноточечный}, Cfr = 0,3, P = 15 \rangle$.

где $(\Delta_{integr})_{\min} = 0,006$; $(\Delta_{integr})_{\max} = 1,776$.

Наилучший результат $(\Delta_{integr} = (\Delta_{integr})_{\min})$ получен при следующих настройках ГА:

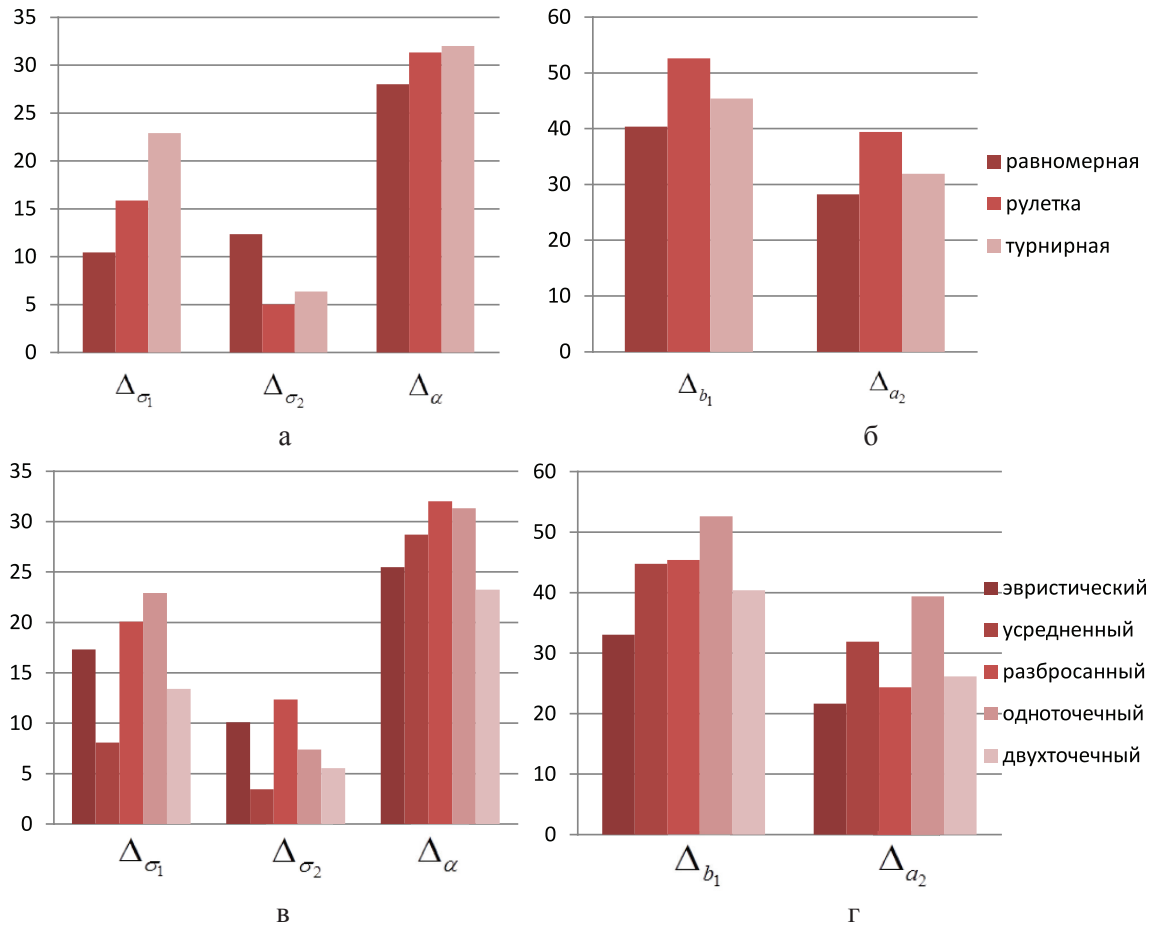


Рис. 3. а – погрешности в разрезе вида селекции – равномерной, на основе рулетки и турнирной селекции; б – погрешности в разрезе вида кроссовера – эвристический, усредненный, разбросанный, одноточечный и двухточечный; в – погрешности в разрезе доли кроссовера – 0,3; 0,6 и 0,9; г – погрешности в разрезе размера популяции – 5, 15 и 30 особей

Соответствующие параметры распределений представлены в таблице.

Теоретическая и экспериментальная плотности распределения представлены на рис. 4.

Наилучший и наихудший интегральные показатели

Параметры	$\tilde{\mu}_1$	$\tilde{\sigma}_1$	$\tilde{a}_1$	$\tilde{b}_1$	$\tilde{\mu}_2$	$\tilde{\sigma}_2$	$\tilde{a}_2$	$\tilde{b}_2$	$\tilde{\alpha}$	Δ_{integr}
Теоретические значения	7	2	3	14	14	1	6	18	0,5	–
Наихудший результат	7,225	1,869	3,023	6,109	14,014	0,956	9,014	17,977	0,813	1,776
Наилучший результат	7,225	2,113	3,023	15,203	14,014	1,051	5,269	17,977	0,474	0,006

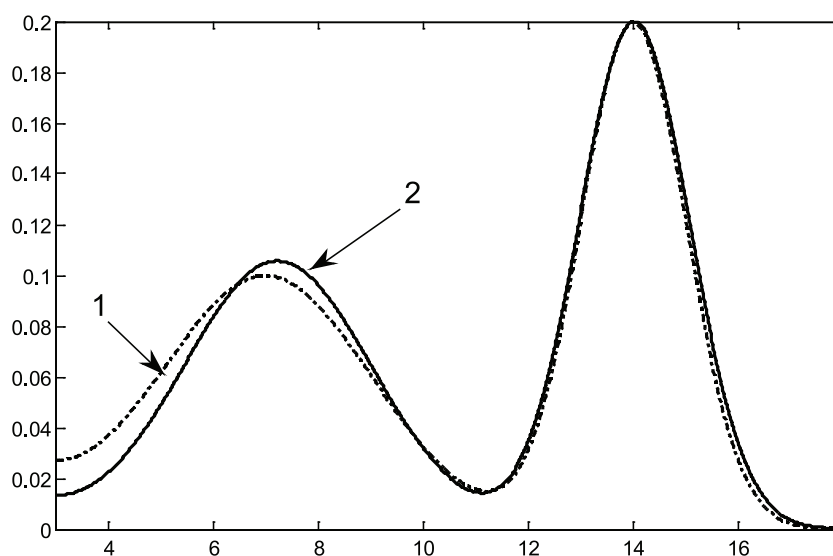


Рис. 4. 1 – Теоретическая и экспериментальная ($\Delta_{integr} = (\Delta_{integr})_{min}$) плотности: теоретическая плотность распределения; 2 – экспериментальная плотность распределения

Таким образом, полученные результаты позволяют сделать вывод о том, что предложенная методика в целом обеспечивает вполне приемлемый результат и ее можно использовать для оценивания параметров двумодальных распределений вида (3).

Выводы

Анализ результатов совместного применения аппроксимации Розенблатта – Парзена, метода мнимых источников и генетических алгоритмов в задаче оценивания значений параметров распределений случайных последовательностей двумодальными законами распределения вида (3), относящегося к классу 9-ти параметрических распределений, позволяет сделать следующие выводы:

1. При использовании генетических алгоритмов для нахождения одновременно всех параметров изучаемого распределения погрешности оценок превышают 50%, что является следствием высокой размерности решаемой задачи.

2. Предложена методика нахождения параметров изучаемого двумодального распределения случайных последовательностей, основанная на совместном использовании аппроксимации Розенблатта – Парзена и ГА, и получено подтверждение ее работоспособности.

3. Получены оценки точности нахождения параметров распределения, а также ин-

тегрального показателя, характеризующего в целом качество оценки плотности распределения случайной последовательности с изученным законом распределения.

Список литературы

1. Крамер Г. Математические методы статистики. – М.: Мир, 1975. – 648 с.
2. Сызранцев В.Н. Расчет прочностной надежности изделий на основе методов непараметрической статистики / В.Н. Сызранцев, Я.П. Невелев, С.Л. Голофаст. – Новосибирск: Наука, 2008. – 218 с.
3. Поршнев С.В. Теория и алгоритмы аппроксимации эмпирических зависимостей и распределений / Е.В. Овечкина, В.Е. Каплан. – Екатеринбург: УрО РАН, 2006. – 166 с.
4. Симахин В.А. Робастные непараметрические оценки: адаптивные оценки взвешенного максимального правдоподобия в условиях статистической априорной неопределенности // Saarbrücken. – Germany: LAP LAMBERT Academic Publishing Gmb H&Co. KG, 2011. – 292 с.
5. Поршнев С.В., Копосов А.С. Использование аппроксимации Розенблатта – Парзена для восстановления функции распределения дискретной случайной величины // В мире научных открытий. – 2013. – № 10(46).
6. Поршнев С.В., Копосов А.С. Использование аппроксимации Розенблатта – Парзена для восстановления функции распределения непрерывной случайной величины с ограниченным одномодальным законом распределения // Политематический сетевой электронный научный журнал Кубанского государственного аграрного университета (Научный журнал КубГАУ) [Электронный ресурс]. – Краснодар: КубГАУ, 2013. – № 08(092). – IDA [article ID]: 0921308076. Режим доступа: <http://ej.kubagro.ru/2013/08/pdf/76.pdf>.
7. Гладков Л.А., Курейчик В.В., Курейчик В.М. Генетические алгоритмы / под ред. В.М. Курейчика. – 2-е изд., исправл. и доп. – М.: ФИЗМАТЛИТ, 2010. – 368 с.

8. Rozenblatt M. Remarks on some nonparametric estimates of density function // Ann. Math. Statist. – 1956. – № 27. – P. 832–835.
9. Parzen E. On estimation of probability density function and mode // Ann. Math. Statist. – 33, 3, 162. – P. 1065–1076.

References

1. Kramer G. Matematicheskie metody statistiki. M.: Mir, 1975. 648 p.
2. Syzrancev V.N. Raschet prochnostnoy nadezhnosti izdeliy na osnove metodov neparametricheskoy statistiki. V.N. Syzrancev, J.P. Nevelev, S.L. Golofast. Novosibirsk: Nauka, 2008. 218 p.
3. Porshnev S.V. Teoriya i algoritmy approksimatsii empiricheskikh zavisimostey i raspredeleniy. E.V. Ovechkina, V.E. Kaplan. Ekaterinburg: UrO RAN, 2006. 166 p.
4. Simahin V.A. Robastnye neparametricheskie ocenki: adaptivnye ocenki vzveshennogo maksimal'nogo pravdopodobiya v usloviyah statisticheskoy apriornoj neopredelennosti. V.A. Simahin. Saarbrücken, Germany: LAP LAMBERT Academic Publishing GmbH & Co. KG, 2011. 292 p.
5. Porshnev S.V., Koposov A.S. Ispol'zovanie approksimatsii Rozenblatta-Parzena dlya vosstanovleniya funktsii raspredeleniya diskretnoy sluchainoy velichiny. V mire nauchnykh otkrytiy. 2013 no. 10(46).
6. Porshnev S.V., Koposov A.S. Ispol'zovanie approksimatsii Rozenblatta-Parzena dlya vosstanovleniya funktsii raspredeleniya nepreryvnoy sluchainoy velichiny s ograniченным odnomodal'nym zakonom raspredeleniya. Politem-

aticheskii setevoy elektronnyy nauchnyy zhurnal Kubanskogo gosudarstvennogo agrarnogo universiteta (Nauchnyy zhurnal KubGAU) [Elektronnyy resurs]. Krasnodar: KubGAU, 2013. no. 08(092). IDA [article ID]: 0921308076. <http://ej.kubagro.ru/2013/08/pdf/76.pdf>.

7. Gladkov L.A., Kurejchik V.V., Kurejchik V.M. Geneticheskie algoritmy. Pod red. V.M. Kurejchika. 2-e izd., ispravl. i dop. M.: FIZMATLIT, 2010. 368 p.

8. Rozenblatt M. Remarks on some nonparametric estimates of density function. Ann. Math. Statist., 1956, 27, 832–835 p.

9. Parzen E. On estimation of probability density function and mode. Ann. Math. Statist., 33, 3, 162, 1065–1076 p.

Рецензенты:

Кубланов В.С., д.т.н., доцент, профессор кафедры радиоэлектроники информационных систем, ГАОУ ВПО «Уральский федеральный университет им. первого Президента России Б.Н. Ельцина», г. Екатеринбург;

Доросинский Л.Г., д.т.н., профессор, заведующий кафедрой информационных технологий, ГАОУ ВПО «Уральский федеральный университет им. первого Президента России Б.Н. Ельцина», г. Екатеринбург.

Работа поступила в редакцию 28.05.2014.