УДК 519.24

СРАВНЕНИЕ ЭМПИРИЧЕСКОЙ И ПРЕДЛАГАЕМОЙ ФУНКЦИЙ РАСПРЕДЕЛЕНИЯ СЛУЧАЙНОЙ ВЕЛИЧИНЫ НА ОСНОВЕ НЕПАРАМЕТРИЧЕСКОГО КЛАССИФИКАТОРА*

А. В. Лапко[1], **В. А. Лапко**[1, 2]

[1] *Учреждение Российской академии наук*
Институт вычислительного моделирования Сибирского отделения РАН,
660036, г. Красноярск, Академгородок, 50, стр. 44
[2] *Государственное образовательное учреждение высшего профессионального образования*
«Сибирский государственный аэрокосмический университет
им. академика М. Ф. Решетнёва»,
660014, г. Красноярск, просп. «Красноярский рабочий», 31
E-mail: lapko@icm.krasn.ru

Рассматривается возможность использования непараметрических алгоритмов распознавания образов в задаче сравнения эмпирической и теоретической функций распределения случайных величин. Проводится анализ результатов вычислительных экспериментов.

Ключевые слова: непараметрическая статистика, распознавание образов, проверка статистических гипотез, распределение случайных величин, критерий Колмогорова.

Введение. Непараметрические алгоритмы распознавания образов, соответствующие критерию максимального правдоподобия, являются эффективным средством при сравнении эмпирических функций распределения случайных величин [1]. Разработанная на их основе методика позволяет обойти трудно формализуемую проблему разбиения области возможных значений случайной величины на интервалы, которая свойственна критерию согласия Пирсона. Предлагаемая методика при проверке гипотезы о тождественности эмпирических законов распределения одномерных случайных величин имеет сопоставимые результаты с критерием Смирнова [2]. Её перспективность заключается в возможности развития на новый класс задач проверки гипотез о распределениях и обобщении на многомерные случайные величины.

Цель данной работы — показать возможность распространения полученных результатов на решение задачи сравнения эмпирической и теоретической функций распределения одномерных случайных величин.

Критерий Колмогорова. Пусть $F_2(x)$ — известная функция распределения одномерной случайной величины x, предполагаемая непрерывной. Имеется реализация $V_1 = (x^i,\ i = \overline{1, n_1})$ случайной величины, которая определяет эмпирическое распределение $\bar{F}_1(x)$. Необходимо проверить гипотезу

$$H_0\colon F_2(x) \equiv F_1(x)$$

о тождественности законов распределения.

*Работа выполнена при поддержке Федеральной целевой программы «Научные и научно-педагогические кадры инновационной России» на 2009–2013 гг. (ГК № 02.740.11.0621).

Для проверки статистической гипотезы H_0 на основе критерия Колмогорова оценим по выборке V_1 функцию распределения $F_1(x)$ случайной величины x:

$$\bar{F}_1(x) = \frac{1}{n_1}\sum_{i=1}^{n_1} 1(x - x^i),$$

где

$$1(x - x^i) = \begin{cases} 0, & \text{если } x - x^i < 0, \\ 1, & \text{если } x - x^i \geq 0. \end{cases}$$

Анализируя значения эмпирической $\bar{F}_1(x)$ и теоретической $F_2(x)$ функций распределения, определим максимальное расхождение между ними:

$$\bar{D}_{12} = \max_x \left|\bar{F}_1(x) - F_2(x)\right|.$$

В соответствии с критерием Колмогорова [3] сравним полученное максимальное расхождение $\bar{D}_{12}$ с пороговым

$$D_\alpha = \sqrt{-\ln(\alpha/2)/(2n_1)}, \tag{1}$$

где α — принятый уровень доверия (риск отвергнуть гипотезу H_0, например, $\alpha = 0{,}05$).

Если выполняется условие $\bar{D}_{12} < D_\alpha$, то гипотеза H_0 справедлива, иначе анализируемые законы распределения различаются.

Непараметрический алгоритм распознавания образов в задаче проверки гипотезы $\boldsymbol{H_0}$. Известно, что если при решении двуальтернативной задачи распознавания образов вероятность ошибки классификации $\rho = 0{,}5$, то законы распределения случайных величин в области определения классов совпадают. Поэтому появляется возможность перехода от задачи сравнения законов распределения случайных величин к проверке гипотезы $\bar{H}_0$ о равенстве вероятности ошибки распознавания образов значению 0,5.

На основании априорной информации осуществим синтез непараметрического алгоритма распознавания образов, соответствующего критерию максимального правдоподобия [4],

$$\bar{m}(x)\colon \begin{cases} x \in \Omega_1, & \text{если } \bar{f}_{12}(x, c) \leq 0, \\ x \in \Omega_2, & \text{если } \bar{f}_{12}(x, c) > 0, \end{cases} \tag{2}$$

где

$$\bar{f}_{12}(x, c) = p_2(x) - \bar{p}_1(x, c)$$

— непараметрическая оценка уравнения разделяющей поверхности между классами Ω_1, Ω_2; $p_2(x)$, $\bar{p}_1(x, c)$ — плотность вероятности распределения x в классе Ω_2 и оценка плотности вероятности $x \in \Omega_1$. При формировании $\bar{f}_{12}(x, c)$ используем непараметрическую оценку плотности вероятности одномерной случайной величины типа Розенблатта — Парзена [5]

$$\bar{p}_1(x, c) = \frac{1}{n_1 c}\sum_{i=1}^{n_1} \Phi\Big(\frac{x - x^i}{c}\Big), \tag{3}$$

восстанавливаемой по статистическим данным $V_1 = (x^i,\ i = \overline{1, n_1})$. Ядерные функции $\Phi(u)$ в статистике (3) удовлетворяют следующим условиям:

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

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

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

а значения их коэффициентов размытости c убывают с ростом количества n_1 элементов выборки V_1. Бесконечные пределы интегрирования в приведённых условиях опускаются.

Выбор оптимального значения $\bar{c}$ коэффициента размытости непараметрического решающего правила $\bar{m}(x)$ осуществляется из условия минимума оценки вероятности ошибки распознавания образов

$$\bar{\rho}(c) = \frac{1}{n_1}\sum_{t=1}^{n_1} 1(\sigma(t),\ \bar{\sigma}(t)),$$

где индикаторная функция

$$1(\sigma(t),\ \bar{\sigma}(t)) = \begin{cases} 0, & \text{если}\,\sigma(t) = \bar{\sigma}(t), \\ 1, & \text{если}\,\sigma(t) \ne \bar{\sigma}(t). \end{cases}$$

Здесь $\sigma(t)$ — «указание» о принадлежности наблюдения x^t из выборки V_1 к классу Ω_1. При вычислении $\bar{\rho}(c)$ «решение» $\bar{\sigma}(t)$ алгоритма (2) об отнесении наблюдения x^t к тому или иному классу определяется в соответствии со знаком статистики

$$\bar{f}_{12}(x^t) = p_2(x^t) - \frac{1}{n_1 c}\sum_{\substack{i=1 \\ i \ne t}}^{n_1} \Phi\Big(\frac{x^t - x^i}{c}\Big),$$

т. е. ситуация x^t, которая подаётся на контроль, исключается из процесса обучения непараметрического алгоритма (2).

В соответствии с критерием Колмогорова проверим гипотезу $\bar{H}_0$: $\bar{\rho}(c) = 0{,}5$. Для этого сравним его пороговое значение (1) с отклонением $\bar{D}_{12} = |0{,}5 - \bar{\rho}(c)|$ при вероятности α отвергнуть правильную гипотезу $\bar{H}_0$.

Если выполняется соотношение $\bar{D}_{12} < D_\alpha$, то гипотеза $\bar{H}_0$ справедлива, иначе она отвергается.

Анализ результатов вычислительных экспериментов. Сравнение эффективности предложенной методики проверки гипотезы о распределениях случайных величин, критериев Колмогорова и Пирсона проводилось по данным вычислительных экспериментов. Последовательность случайных наблюдений $V_1 = (x^i,\ i = \overline{1, n_1})$ формировалась на основе датчиков случайных величин с равномерным $x^i = \varepsilon^i$ и нормальным $x^i = 0{,}5 + 0{,}15\Big(\sum\limits_{j=1}^{12}\varepsilon^j - 6\Big)$, $i = \overline{1, n}$, законами распределения. Случайные величины ε с равномерным законом распределения определены на интервале $[0, 1]$. При их формировании использовался стандартный датчик псевдослучайных величин среды визуального программирования Delphi.

Вычислительные эксперименты при фиксированных условиях исследования осуществлялись $N = 100$ раз. При априори тождественных законах распределения случайных ве-

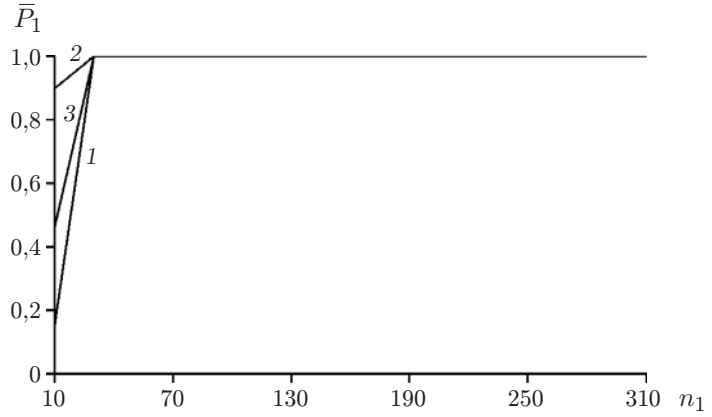

Рис. 1. Зависимости оценок $\bar{P}_1$ вероятностей отклонения гипотезы H_0 от объёма n_1 выборки V_1 при сравнении равномерного и нормального законов распределения случайных величин. Кривые *1–3* получены при использовании критериев Колмогорова, Пирсона и исследуемой методики соответственно

личин по полученным результатам оценивалась вероятность P_2 выполнения гипотезы H_0. Если сравниваемые законы распределения отличались, то оценивалась вероятность P_1 отклонения гипотезы H_0. Риск α отвергнуть гипотезу H_0 принимался равным 0,05.

При использовании критерия согласия Пирсона количество интервалов r, на которые разбивалась область изменения случайной величины, определялось в соответствии с формулой Старджесса. Так как теоретические законы распределения были заданы априори, то количество степеней свободы принималось равным $r - 1$ [3].

При синтезе непараметрического классификатора $\bar{m}(x)$ использовались параболические ядерные функции Епанечникова [6].

Результаты вычислительных экспериментов, когда сравниваемые законы распределений случайных величин разные, представлены на рис. 1. При $n_1 > 20$ рассматриваемые критерии безошибочно отклоняют гипотезу H_0. В интервале малых значений $n_1 < 20$ применение сравниваемых критериев приводит к неудовлетворительным результатам, что, возможно, зависит от качества используемого датчика случайных величин.

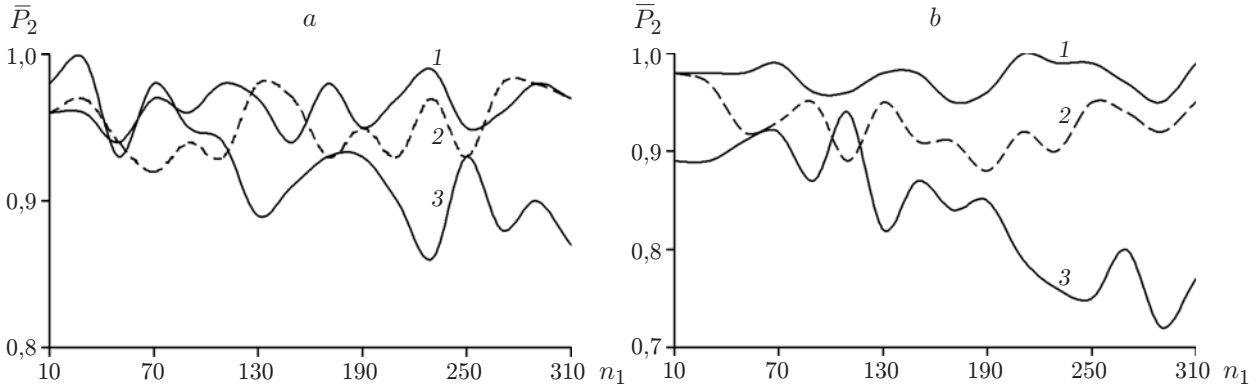

Рис. 2. Зависимости оценок $\bar{P}_2$ вероятностей справедливости гипотезы H_0 от объёма n_1 выборки V_1 в условиях априори одинаковых законов распределения: равномерные (a), нормальные (b). Обозначения кривых соответствуют рис. 1

Если априори законы распределения случайных величин тождественны, то для достаточно широкого интервала изменения объёма n_1 анализируемой выборки оценки вероятности справедливости гипотезы H_0 при использовании критериев Колмогорова, Пирсона и предлагаемой методики сопоставимы (рис. 2, *a*). При увеличении $n_1 > 200$ наблюдается снижение эффективности исследуемой методики, что особенно характерно при сравнении нормальных законов случайных величин (рис. 2, *b*).

Заключение. В представленной работе показана возможность применения непараметрического алгоритма распознавания образов, соответствующего критерию максимального правдоподобия, в задаче сравнения эмпирической и теоретической функций распределения случайных величин. Существуют условия, когда использование предлагаемой методики и критериев Колмогорова, Пирсона приводит к сопоставимым результатам.

Перспективность такой методики заключается в возможности её обобщения на задачу проверки гипотез о распределениях многомерных случайных величин с обходом проблемы разбиения области их значений на интервалы.

СПИСОК ЛИТЕРАТУРЫ

1. **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.
2. **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.
3. **Пугачев В. С.** Теория вероятностей и математическая статистика. М.: Наука, 1979. 496 с.
4. **Лапко А. В., Лапко В. А., Соколов М. И., Ченцов С. В.** Непараметрические системы классификации. Новосибирск: Наука, 2000. 240 с.
5. **Parzen E.** On estimation of a probability density function and mode // Ann. Math. Statist. 1962. **33**, N 3. P. 1065–1076.
6. **Епанечников В. А.** Непараметрическая оценка многомерной плотности вероятности // Теория вероятностей и ее применения. 1969. **14**, вып. 1. С. 156–161.

Поступила в редакцию 8 февраля 2011 г.