УДК 681.513

НЕПАРАМЕТРИЧЕСКИЕ АЛГОРИТМЫ РАСПОЗНАВАНИЯ ОБРАЗОВ В ЗАДАЧЕ ПРОВЕРКИ СТАТИСТИЧЕСКОЙ ГИПОТЕЗЫ О ТОЖДЕСТВЕННОСТИ ДВУХ ЗАКОНОВ РАСПРЕДЕЛЕНИЯ СЛУЧАЙНЫХ ВЕЛИЧИН*

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

Предлагаются методики проверки гипотез о распределениях случайных величин, основанные на использовании непараметрических алгоритмов распознавания образов. По результатам вычислительных экспериментов проводится их сравнение с критерием Колмогорова — Смирнова.

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

Введение. Задачи проверки гипотез о распределениях случайных величин являются одними из основных в теории вероятностей и математической статистике. Для их решения широко используется критерий согласия Пирсона, который не зависит от распределений случайных величин и их размерности. Его применение позволяет, например, проверять гипотезы о тождественности эмпирического и гипотетического законов распределений, а также гипотезы о совпадении распределений в двух выборках случайных величин [1]. Однако методика формирования критерия Пирсона содержит трудно формализуемый этап разбиения области возможных значений случайной величины на многомерные интервалы. Этот этап отсутствует в критерии Колмогорова — Смирнова, который обоснован и используется при проверке гипотез о распределениях одномерных случайных величин [2].

Цель данной работы состоит в обосновании возможности применения непараметрических алгоритмов распознавания образов, основанных на ядерных оценках плотности вероятности типа Розенблатта — Парзена [3], в задаче проверки статистической гипотезы о тождественности двух эмпирических законов распределения одномерных случайных величин. Решение указанной проблемы создаёт основу обобщения предлагаемого подхода на сравнение распределений многомерных случайных величин.

Критерий Колмогорова — Смирнова. 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 этих генеральных совокупностей, проверить либо опровергнуть гипотезу о тождественности законов распределения $H_0\colon P(X_1) \equiv P(X_2)$.

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

Методика проверки статистической гипотезы H_0 на основе критерия Колмогорова — Смирнова сводится к выполнению следующих действий:

1. По независимым выборкам V_1, V_2 построить оценки функций распределения

$$\bar{P}_j(x) = \frac{1}{n_j}\sum_{i=1}^{n_j} 1(x - x^i),$$

где

$$1(x - x^i) = \begin{cases} 0, & \text{если } x - x^i < 0, \\ 1, & \text{если } x - x^i \geq 0, \end{cases} \qquad j = 1,\ 2.$$

2. Найти максимальное расхождение между эмпирическими функциями распределения:

$$\bar{D}_{12} = \max_x |\bar{P}_1(x) - \bar{P}_2(x)|.$$

3. В соответствии с критерием Колмогорова — Смирнова [2, 4] сравнить полученное максимальное расхождение $\bar{D}_{12}$ с пороговым

$$D_\alpha = \sqrt{-\ln\frac{\alpha}{2}\Big(\frac{1}{n_1} + \frac{1}{n_2}\Big)\Big/2}, \tag{1}$$

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 эмпирические законы распределения различаются.

Методика 1 проверки гипотезы о распределениях. Известно, что если при решении двухальтернативной задачи распознавания образов вероятность ошибки классификации равна 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 статистической оценки вероятности ошибки распознавания образов значению 0,5.

Сформируем на основе независимых последовательностей случайных величин V_1 и V_2 обучающую выборку $V = (x^i,\ \sigma(i),\ i = \overline{1,n})$ для решения задачи распознавания образов, где $n = n_1 + n_2$; $\sigma(i)$ — указание о принадлежности значения x^i к тому либо иному классу Ω_1, Ω_2. При этом полагается, что элементы множеств V_1 и V_2 принадлежат классам Ω_1, Ω_2 соответственно. На этой основе осуществим синтез непараметрического алгоритма распознавания образов, соответствующего критерию максимального правдоподобия [5]

$$\bar{m}(x)\colon \begin{cases} x \in \Omega_1, & \text{если } \bar{f}_{12}(x, c_1, c_2) \leq 0, \\ x \in \Omega_2, & \text{если } \bar{f}_{12}(x, c_1, c_2) > 0, \end{cases} \tag{2}$$

где

$$\bar{f}_{12}(x, c_1, c_2) = \bar{p}_2(x, c_2) - \bar{p}_1(x, c_1). \tag{3}$$

В статистике (3)

$$\bar{p}_j(x, c_j) = \frac{1}{n_j\, c_j}\sum_{i \in I_j} \Phi\Big(\frac{x - x^i}{c_j}\Big) \tag{4}$$

— непараметрическая оценка плотности вероятности распределения случайной величины x в j-м классе, которая восстанавливается по наблюдениям $x^i \in V_j$, $j = 1, 2$; I_j — множество номеров значений $x^i \in V_j$ в обучающей выборке V.

Ядерные функции в статистике (4) удовлетворяют 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_j убывают с ростом количества n_j элементов множества V_j, $j = 1, 2$.

В рассматриваемой методике проверки статистических гипотез о тождественности распределений выбор оптимальных $\bar{c}_j$ коэффициентов размытости ядерных функций предлагается осуществлять из условия максимума функции правдоподобия

$$L_j(c_j) = \prod_{t \in I_j} \bar{p}_j(x^t, c_j),$$

где

$$\bar{p}_j(x^t, c_j) = \frac{1}{(n_j - 1)c_j} \sum_{\substack{i \in I_j \\ i \ne t}} \Phi\Big(\frac{x^t - x^i}{c_j}\Big), \quad j = 1, 2.$$

Непараметрическое решающее правило (2) при оптимальных значениях $\bar{c}_j$, $j = 1, 2$, является основой синтеза критерия проверки статистической гипотезы H_0. Для этого рассчитаем оценку вероятности ошибки распознавания образов в режиме «скользящего экзамена»:

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

где индикаторная функция

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

$\bar{\sigma}(t)$ — «решение» о принадлежности значений x^t к тому либо иному классу Ω_1, Ω_2, полученное в соответствии с алгоритмом распознавания образов (2). При формировании решения $\bar{\sigma}(t)$ в режиме скользящего экзамена номер ситуации $x^t \in \Omega_j$ исключается из множества I_j в непараметрической статистике $\bar{p}_j(x, c_j)$ (4).

Проверим гипотезу $\bar{H}_0$: $\bar{\rho} = 0{,}5$ в соответствии с критерием Колмогорова. Для этого сравним его пороговое значение

$$\bar{D}_\alpha = \sqrt{-\ln\frac{\alpha}{2}\Big(\frac{1}{n_1 + n_2}\Big)\Big/2} \tag{5}$$

с отклонением $\bar{D}_{12} = |0{,}5 - \bar{\rho}|$ при вероятности α отвергнуть правильную гипотезу $\bar{H}_0$.

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

Методика 2 проверки гипотезы о распределениях. В отличие от вышерассмотренной методики 1 в непараметрическом алгоритме распознавания образов

$$\bar{\bar{m}}(x)\colon \begin{cases} x \in \Omega_1, & \text{если } \bar{f}_{12}(x) \le 0, \\ x \in \Omega_2, & \text{если } \bar{f}_{12}(x) > 0, \end{cases} \tag{6}$$

оценка уравнения разделяющей поверхности определяется выражением

$$\tilde{f}_{12}(x) = (nc)^{-1} \sum_{i=1}^{n} \sigma_1(i) \Phi\Big(\frac{x - x^i}{c}\Big),$$

где

$$\sigma_1(i) = \begin{cases} -\bar{P}_1^{-1}, & \text{если } x^i \in \Omega_1, \\ \bar{P}_2^{-1}, & \text{если } x^i \in \Omega_2, \end{cases}$$

$\bar{P}_j = n_j/n$ — оценка априорной вероятности принадлежности ситуаций обучающей выборки к классу Ω_j, $j = 1, 2$.

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

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

При вычислении $\bar{\bar{\rho}}$ решение $\bar{\sigma}(t)$ алгоритма (6) определяется в соответствии со знаком статистики

$$\tilde{f}_{12}(x^t) = (n\bar{c})^{-1} \sum_{\substack{i=1 \\ i \ne t}}^{n} \sigma_1(i) \Phi\Big(\frac{x^t - x^i}{\bar{c}}\Big),$$

т. е. ситуация x^t, которая подаётся на контроль, исключается из процесса обучения.

Следуя предложенной методике, сравним отклонение $\bar{\bar{D}}_{12} = |0{,}5 - \bar{\bar{\rho}}|$ с пороговым значением (5). Гипотеза $\bar{H}_0$ справедлива, если выполняется неравенство $\bar{D}_{12} < \bar{D}_\alpha$, иначе она отвергается.

Анализ результатов вычислительных экспериментов. Проводилось сравнение эффективности предложенных методик проверки гипотез о распределениях случайных величин и критерия Колмогорова — Смирнова по данным вычислительных экспериментов. Последовательности случайных наблюдений $V_1 = (x^i,\, i = \overline{1,\, n_1})$ и $V_2 = (x^i,\, i = \overline{1,\, n_2})$ формировались на основе датчиков случайных величин с равномерным $x^i = \varepsilon^i$, нормальным $x^i = 0{,}5 + 0{,}15\Big(\sum\limits_{j=1}^{12} \varepsilon^j - 6\Big)$ и показательным $x^i = \sqrt[4]{\varepsilon^i}$, $i = \overline{1,\, n}$, 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.

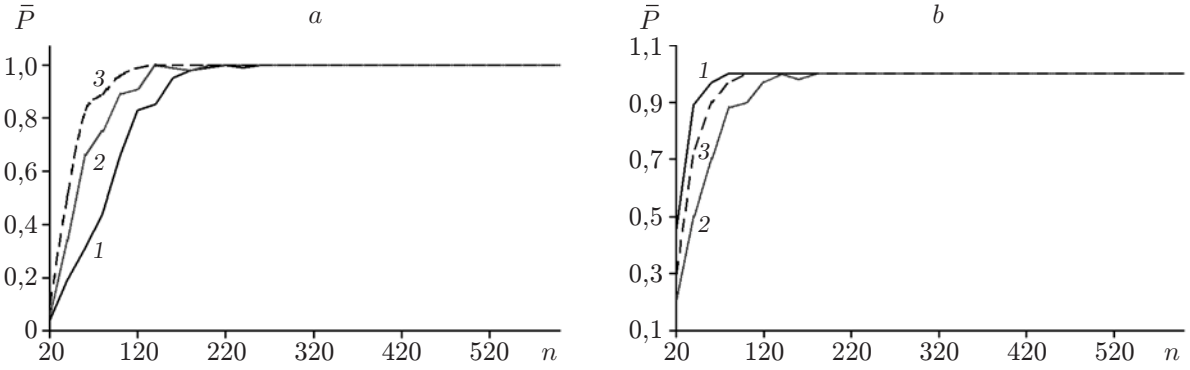

Рис. 1. Зависимости оценок вероятностей отклонения гипотезы H_0 от объёма экспериментальных данных $n = n_1 + n_2$ при $n_1 = n_2$ в условиях, когда сравниваемые законы распределения являются разными: равномерный и нормальный (a), равномерный и показательный (b). Здесь и далее на рисунках кривые *1* получены при использовании критерия Колмогорова — Смирнова, кривые *2* — методики 1, кривые *3* — методики 2

Вычислительные эксперименты при фиксированных условиях осуществлялись $N = 100$ раз. По полученным результатам при априори тождественных законах распределения случайных величин оценивалась вероятность выполнения гипотезы H_0. Если законы распределения различались, то оценивалась вероятность её отклонения. Риск α отвергнуть гипотезу H_0 принимался равным 0,05.

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

Результаты вычислительных экспериментов при разных сравниваемых законах распределения случайных величин представлены на рис. 1, 2. При $n > 200$ рассматриваемые критерии безошибочно отклоняют гипотезу H_0. Однако при $60 < n < 200$ наблюдается значимое преимущество методик 1 и 2 над критерием Колмогорова — Смирнова при сравнении последовательностей случайных величин с равномерным и нормальным законами распределения (рис. 1, a, 2, a). Отмеченный факт свойствен как условиям равных значений $n_1 = n_2$, так и случаю, когда количество элементов сравниваемых последовательностей разное ($n_1 = 2n_2$). При сравнении последовательностей случайных величин с равномерным и показательным законами распределения эффективность сравниваемых критериев одинакова (рис. 1, b, 2, b). В интервале малых значений ($n < 60$) применение

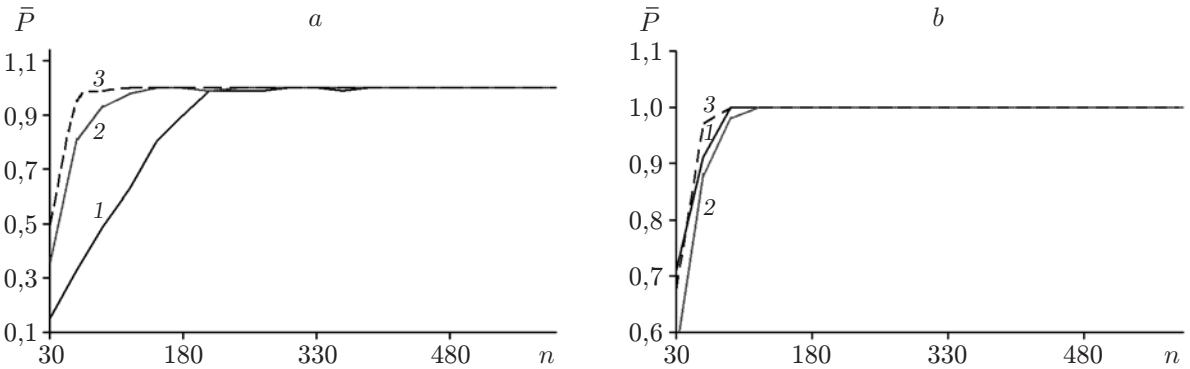

Рис. 2. Зависимости оценок вероятностей отклонения гипотезы H_0 от объёма экспериментальных данных $n = n_1 + n_2$ при $n_1 = 2n_2$. Условия эксперимента, как на рис. 1

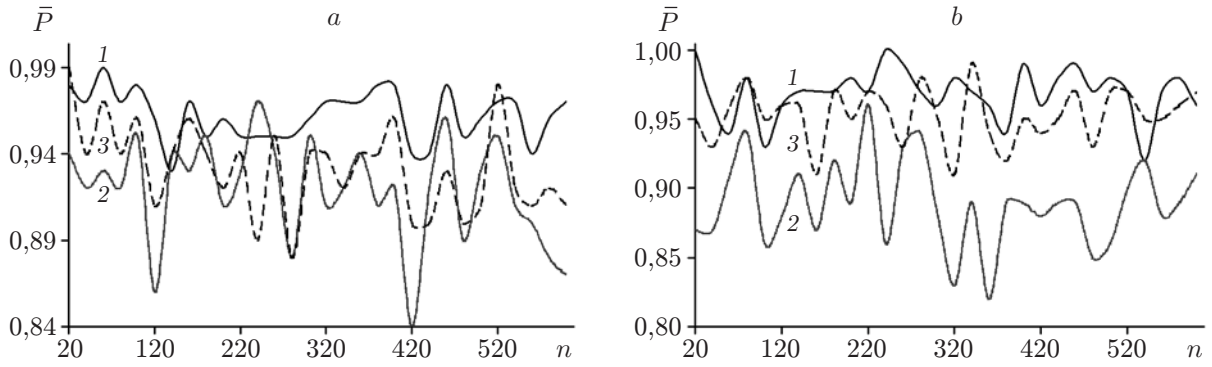

Рис. 3. Зависимости оценок вероятностей справедливости гипотезы H_0 от объёма экспериментальных данных $n = n_1 + n_2$ при $n_1 = n_2$ в условиях априори одинаковых законов распределения: равномерные (a), нормальные (b)

сравниваемых критериев приводит к неудовлетворительным результатам, что, возможно, зависит от качества используемого датчика случайных величин.

Если априори законы распределения случайных величин тождественны при $n_1 = n_2$ на всём интервале изменения $n = n_1 + n_2$, оценки вероятности справедливости гипотезы H_0 при использовании критерия Колмогорова — Смирнова и методики 2 достоверно не отличаются (рис. 3).

При различных объёмах случайных последовательностей $n_1 = 2n_2$ установлено снижение эффективности методики 2 по сравнению с критерием Колмогорова — Смирнова (рис. 4). Данный факт согласуется с результатами исследований работы [7], где показано значительное снижение аппроксимационных свойств непараметрической оценки уравнения разделяющей поверхности при увеличении степени неравномерности распределения элементов обучающей выборки между классами. Повышение эффективности методики 2 в данных условиях возможно на основе построения семейства непараметрических уравнений разделяющей поверхности при равновероятном распределении элементов обучающей выборки между классами с последующим их объединением в коллективе решающих функций. Синтез каждой составляющей коллектива при $n_1 > n_2$ осуществляется по обучающей выборке объёма $n = 2n_2$, в которой множество элементов второго класса образует сравниваемая последовательность V_2, а элементы первого класса формируются случайным образом из последовательности V_1.

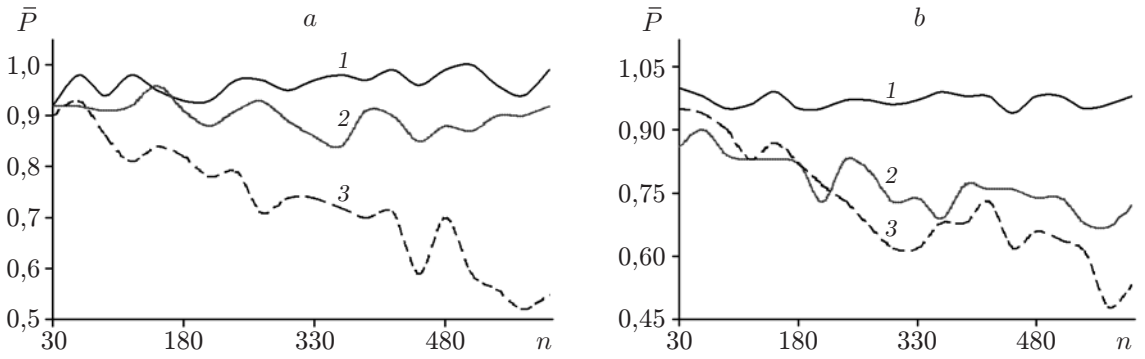

Рис. 4. Зависимости оценок вероятностей справедливости гипотезы H_0 от объёма экспериментальных данных $n = n_1 + n_2$ при $n_1 = 2n_2$. Условия эксперимента, как на рис. 3

Заключение. В данной работе показана возможность применения непараметрических алгоритмов распознавания образов, соответствующих критерию максимального правдоподобия, в задаче проверки статистических гипотез о распределениях случайных величин. Существуют условия, когда использование предлагаемых методик и критерия Колмогорова — Смирнова приводит к сопоставимым результатам. К ним относятся задачи проверки гипотез при разных законах распределения случайных величин и при одинаковых законах распределения, когда отношение объёмов сравниваемых последовательностей принадлежит интервалу [0,8; 1,2].

Перспективность предлагаемых методик заключается в возможности их обобщения на задачу проверки гипотез о тождественности законов распределения многомерных случайных величин.

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. ун-та. 1930. **2**, № 2. С. 3–14.
3. **Parzen E.** On estimation of a probability density function and mode // Ann. Math. Statist. 1962. **33**, N 3. P. 1065–1076.
4. **Шаракшанэ А. С., Железнов И. Г., Ивницкий В. А.** Сложные системы. М.: Высш. шк., 1977. 247 с.
5. **Лапко А. В., Лапко В. А., Соколов М. И., Ченцов С. В.** Непараметрические системы классификации. Новосибирск: Наука, 2000. 240 с.
6. **Епанечников В. А.** Непараметрическая оценка многомерной плотности вероятности // Теория вероятностей и её применения. 1969. **14**, вып. 1. С. 156–161.
7. **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.

Поступила в редакцию 25 июня 2010 г.