УДК 519.7

ИССЛЕДОВАНИЕ МЕТОДИКИ ПРОВЕРКИ ГИПОТЕЗЫ О НЕЗАВИСИМОСТИ ДВУХМЕРНЫХ СЛУЧАЙНЫХ ВЕЛИЧИН С ИСПОЛЬЗОВАНИЕМ НЕПАРАМЕТРИЧЕСКОГО КЛАССИФИКАТОРА

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

[1] *Институт вычислительного моделирования СО РАН,*
660036, г. Красноярск, Академгородок, 50, стр. 44
[2] *Сибирский государственный университет науки и технологий им. академика М. Ф. Решетнева,*
660037, г. Красноярск, просп. им. газеты «Красноярский рабочий», 31
E-mail: lapko@icm.krasn.ru

Исследуются свойства методики проверки гипотезы о независимости случайных величин, основанной на использовании непараметрического алгоритма распознавания образов, соответствующего критерию максимального правдоподобия. Оценивание законов распределения в классах осуществляется по исходным статистическим данным в предположении независимости и зависимости сравниваемых случайных величин. В этих условиях вычисляются оценки вероятностей ошибок распознавания образов в классах. По минимальному их значению принимается решение о независимости либо зависимости случайных величин. Применение предлагаемой методики позволяет обойти проблему декомпозиции области значений случайных величин на многомерные интервалы. Эффективность предлагаемой методики при усложнении зависимости между случайными величинами и изменении объёма исходных статистических данных исследуется методом вычислительных экспериментов.

Ключевые слова: проверка гипотезы о независимости случайных величин, двухмерные случайные величины, непараметрический алгоритм распознавания образов, ядерная оценка плотности вероятности, критерий максимального правдоподобия, критерий Пирсона, зависимые случайные величины.

DOI: 10.15372/AUT20210610

Введение. Информация о независимости либо зависимости случайных величин является необходимой при выборе значимых переменных и построении в пространстве их значений эффективных алгоритмов принятия решений. Традиционная методика проверки гипотезы о независимости случайных величин основана на использовании критерия Пирсона, которая содержит трудно формализуемый этап разбиения области значений случайных величин на многомерные интервалы [1]. Поэтому актуальной является задача разработки новой методики проверки гипотезы о независимости случайных величин, которая обеспечивает обход проблемы декомпозиции области их значений на многомерные интервалы.

В работах [2–4] предложено решение подобной задачи при проверке гипотезы о тождественности законов распределения случайных величин на основе использования непараметрического алгоритма распознавания образов, соответствующего критерию максимального правдоподобия. Показана возможность её замены задачей проверки гипотезы о равенстве ошибки распознавания образов определённому пороговому значению. Обучающая выборка при синтезе непараметрического алгоритма распознавания образов формируется по статистическим данным, характеризующим законы распределения сравниваемых случайных величин. Предложенный подход был развит при проверке гипотезы о независимости случайных величин [5]. В этих условиях по исходным статистическим данным формируется обучающая выборка для решения двухальтернативной задачи распознавания образов.

Каждый класс определяется исходными статистическими данными в предположении их независимости либо зависимости, что проявляется в различии законов распределения случайных величин в классах. Разработанная методика подтверждена результатами вычислительных экспериментов при условии линейной зависимости между случайными величинами.

Цель предлагаемой работы состоит в исследовании путём организации вычислительных экспериментов эффективности методики проверки гипотезы о независимости случайных величин в условиях, когда они характеризуются усложнением функциональной зависимости и различными объёмами статистических данных.

Модификация непараметрического алгоритма распознавания образов, соответствующего критерию максимального правдоподобия. Имеется выборка $V = ((x_1, y_1), \ldots, (x_n, y_n))$ значений двух совместно наблюдаемых случайных величин X и Y в n независимых экспериментах. Требуется проверить гипотезу

$$H_0: \quad \{X \text{ и } Y \text{ независимы}\}. \tag{1}$$

Для проверки гипотезы (1) будем решать двухальтернативную задачу распознавания образов. Под классами Ω_1, Ω_2 понимаются области определения плотностей вероятностей $p(X)p(Y)$, $p(X,Y)$, характеризующих условия независимости и зависимости случайных величин X и Y. Тогда байесовское решающее правило, соответствующее критерию максимального правдоподобия, имеет вид

$$m(X,Y): \qquad \begin{cases} (X,Y) \in \Omega_1, & \text{если } p(X,Y) < p(X)p(Y); \\ (X,Y) \in \Omega_2, & \text{если } p(X,Y) > p(X)p(Y). \end{cases} \tag{2}$$

В отличие от традиционной постановки задачи распознавания образов при синтезе решающего правила (2) априори отсутствует обучающая выборка, содержащая сведения о принадлежности элементов выборки V к тому или иному классу. Эти сведения заменяются на предположения о независимости либо зависимости случайных величин в соответствии с гипотезой (1). В этих условиях по выборке V восстановим плотности вероятностей $p(X,Y)$, $p(X)p(Y)$, используя их непараметрические оценки типа Розенблатта — Парзена [6, 7]:

$$\bar{p}(X,Y) = \frac{1}{nc_1c_2} \sum_{i=1}^{n} \Phi\Big(\frac{X - x^i}{c_1}\Big)\Phi\Big(\frac{Y - y^i}{c_2}\Big), \tag{3}$$

$$\bar{p}(X)\bar{p}(Y) = \frac{1}{n^2c_1c_2} \sum_{i=1}^{n}\sum_{j=1}^{n} \Phi\Big(\frac{X - x^i}{c_1}\Big)\Phi\Big(\frac{Y - y^i}{c_2}\Big). \tag{4}$$

В статистике, например, (3) ядерные функции

$$\Phi(u) = \frac{1}{c_1}\,\Phi\Big(\frac{X - x^i}{c_1}\Big)$$

удовлетворяют следующим условиям:

$$\Phi(u) = \Phi(-u), \qquad 0 \leqslant \Phi(u) < \infty, \qquad \int \Phi(u)\,du = 1,$$

$$\int u^m \Phi(u)\,du < \infty, \qquad 0 \leqslant m < \infty.$$

Здесь и далее бесконечные пределы интегрирования опускаются.

Значения коэффициентов размытости c_1, c_2 ядерных функций убывают с ростом объёма n выборки статистических данных V. С учётом выражений (2)–(4) непараметрическое решающее правило классификации случайных величин (X, Y) запишется как

$$\bar{m}(X,Y): \quad \begin{cases} (X,Y) \in \Omega_1, & \text{если } \bar{p}(X,Y) < \bar{p}(X)\bar{p}(Y); \\ (X,Y) \in \Omega_2, & \text{если } \bar{p}(X,Y) > \bar{p}(X)\bar{p}(Y). \end{cases} \tag{5}$$

Оптимальный коэффициент размытости c_1 ядерных функций, например, непараметрической оценки плотности вероятности $\bar{p}(X)$, будем определять из условия минимума критерия

$$W(c_1) = \int \left(\bar{p}(X) - p(X)\right)^2 dX, \tag{6}$$

который характеризует меру близости между $\bar{p}(X)$ и $p(X)$.

Преобразуем с учётом непараметрической оценки плотности вероятности $\bar{p}(X)$ выражение (6):

$$W(c_1) = \frac{1}{n^2 c_1^2} \sum_{j=1}^{n} \sum_{i=1}^{n} \int \Phi\Big(\frac{X - x^j}{c_1}\Big) \Phi\Big(\frac{X - x^i}{c_1}\Big)\, dX -$$

$$- \frac{2}{n c_1} \sum_{i=1}^{n} \int \Phi\Big(\frac{X - x^i}{c_1}\Big) p(X)\, dX + \int p^2(X)\, dX.$$

Третий член последнего выражения не зависит от c_1, поэтому при минимизации критерия $W(c_1)$ его можно не учитывать. Вид второго слагаемого $b(c_1)$ допускает оценивание статистикой

$$\bar{b}(c_1) = -\frac{2}{n^2 c_1} \sum_{j=1}^{n} \sum_{\substack{i=1 \\ i \neq j}}^{n} \Phi\Big(\frac{x^j - x^i}{c_1}\Big).$$

При выполнении условия $i \neq j$ статистика $\bar{b}(c_1)$ является несмещённой оценкой

$$b(c_1) = -2 \int \bar{p}(X) p(X)\, dX.$$

Тогда оптимальные значения $\bar{c}_1$ будем находить путём минимизации критерия

$$\bar{W}(c_1) = \frac{1}{n^2 c_1^2} \sum_{j=1}^{n} \sum_{i=1}^{n} \int \Phi\Big(\frac{X - x^j}{c_1}\Big) \Phi\Big(\frac{X - x^i}{c_1}\Big)\, dX - \frac{2}{n^2 c_1} \sum_{j=1}^{n} \sum_{\substack{i=1 \\ i \neq j}}^{n} \Phi\Big(\frac{x^j - x^i}{c_1}\Big). \tag{7}$$

Возможность использования критерия (7) для выбора оптимальных коэффициентов размытости в $\bar{p}(X)$ заключается в том, что статистическая оценка $\bar{b}(c_1)$ имеет значительно бóльшую скорость сходимости к $b(c_1)$ с ростом n, чем $\bar{p}(X)$ к $p(X)$. По аналогии с выражением (7) нетрудно определить критерии выбора оптимальных коэффициентов размытости непараметрических статистик $\bar{p}(Y)$, $\bar{p}(X,Y)$ (3).

Впервые подход к оптимизации непараметрической оценки плотности вероятности типа Розенблатта — Парзена по коэффициенту размытости ядерных функций из условия минимума статистической оценки среднего квадратического отклонения, например $\bar{p}(X)$ от $p(X)$, был предложен в [8]. Эта методика позднее была повторена в [9–11] и является актуальной до настоящего времени [12–15]. Исследованы её свойства при использовании ядерных функций, соответствующих нормальному закону [15]. В этих условиях значительно упрощается вычисление критерия оптимизации $\bar{p}(X)$ по значению c_1. Выбор оптимальных значений коэффициентов размытости ядерных функций, соответствующих максимуму функции правдоподобия, рассмотрен в [16, 17].

Оптимизацию непараметрического решающего правила (5) по коэффициентам размытости ядерных функций c_1, c_2 можно упростить, если полагать в статистиках (3), (4) значения $c_1 = c\bar{\sigma}_1$, $c_2 = c\bar{\sigma}_2$. Здесь $\bar{\sigma}_1$, $\bar{\sigma}_2$ — оценки средних квадратических отклонений случайных величин X, Y, вычисляемых по выборке V. Данное утверждение является очевидным, так как большей длине интервалов значений X, Y соответствуют бóльшие коэффициенты размытости ядерных функций. Подобный подход использовался при построении быстрых процедур оптимизации непараметрических оценок плотности вероятности ядерного типа [18–22].

Поэтому появляется возможность оптимизацию непараметрического алгоритма распознавания образов (5) проводить лишь по одному параметру c коэффициентов размытости ядерных функций. Подобный подход использовался при построении быстрых процедур оптимизации непараметрической оценки плотности вероятности [23].

Методика проверки гипотезы о независимости случайных величин. Непараметрический алгоритм распознавания образов (5) основан на проверке соотношений между ядерными оценками плотностей вероятностей $\bar{p}(X)\bar{p}(Y)$ и $\bar{p}(X,Y)$. Для ситуаций первого класса Ω_1, в которых соотношение $\bar{p}(X)\bar{p}(Y) > \bar{p}(X,Y)$, справедливо предположение о независимости случайных величин X, Y. В области определения непараметрической оценки плотности вероятности $\bar{p}(X,Y)$ при выполнении соотношения $\bar{p}(X)\bar{p}(Y) < \bar{p}(X,Y)$ следует зависимость случайных величин. Выполнение гипотезы (1) определяет границу в области значений случайных величин X, Y, разделяющих предположения о независимости либо зависимости X, Y. С этих позиций методика проверки гипотезы о независимости случайных величин предполагает выполнение следующих действий.

1. В соответствии с рекомендациями предыдущего раздела осуществить синтез непараметрического алгоритма распознавания образов (5).

2. Определить оценки вероятностей ошибок распознавания образов $\bar{\rho}_1$, $\bar{\rho}_2$ решающим правилом (5) по исходным статистическим данным V при оптимальных коэффициентах размытости ядерных статистик $\bar{p}(X)\bar{p}(Y)$, $\bar{p}(X,Y)$.

Значения $\bar{\rho}_t$ вычисляются в режиме «скользящего экзамена» по выборке V в предположении, что её элементы принадлежат к классу Ω_t:

$$\bar{\rho}_t = \frac{1}{n}\sum_{j=1}^{n} 1(\delta(j), \bar{\delta}(j)), \qquad t = 1, 2,$$

где $\delta(j)$ — указания типа $(x^j, y^j) \in \Omega_t$; $\bar{\delta}(j)$ — «решение» алгоритма (5) о принадлежности ситуации (x^j, y^j) к одному из классов Ω_t, $t = 1, 2$.

При вычислении $\bar{\rho}_t$ в соответствии с методикой «скользящего экзамена» ситуация (x^j, y^j) из выборки V, которая подаётся на контроль в алгоритм (5), исключается из процесса формирования статистик (3), (4).

Индикаторная функция определяется выражением

$$1(\delta(j), \bar{\delta}(j)) = \begin{cases} 0, & \text{если } \delta(j) = \bar{\delta}(j); \\ 1, & \text{если } \delta(j) \neq \bar{\delta}(j). \end{cases}$$

3. Сравнить значения $\bar{\rho}_1$, $\bar{\rho}_2$ в предположении, что элементы выборки V принадлежат к классам Ω_1, Ω_2 соответственно. Тогда гипотеза H_0 справедлива, если $\bar{\rho}_1 < \bar{\rho}_2$. В противном случае при $\bar{\rho}_2 < \bar{\rho}_1$ случайные величины X и Y являются зависимыми.

При ограниченных объёмах n выборки V возникает задача доверительного оценивания вероятностей ошибок распознавания образов $\bar{\rho}_1$, $\bar{\rho}_2$. Для её решения может применяться традиционная методика доверительного оценивания вероятностей [1] либо критерий Колмогорова — Смирнова [24].

Например, при использовании критерия Колмогорова — Смирнова отклонение $D_{12} = |\bar{\rho}_1 - \bar{\rho}_2|$ сравнивается с пороговым значением

$$D_\beta = \sqrt{-\ln(\beta/2)/n}.$$

Здесь β — вероятность (риск) отвергнуть гипотезу $\bar{H}_0$: $\rho_1 = \rho_2$. Если выполняется соотношение $D_{12} < D_\beta$, то гипотеза $\bar{H}_0$ справедлива и риск её отвергнуть не превышает значения β. При $D_{12} > D_\beta$ гипотеза $\bar{H}_0$ отвергается.

Анализ результатов вычислительных экспериментов. Исследуем эффективность предлагаемой методики от объёма n исходных статистических данных и усложнения зависимостей между случайными величинами.

При организации вычислительных экспериментов случайные величины X, Y принимаются зависимыми. Значения X формировались с помощью датчика с равномерным законом распределения $p(X) = R(0; 1{,}732)$ при математическом ожидании ноль и среднем квадратическом отклонении 1,732.

Значения случайной величины Y формировались с нормальным законом распределения $N(\varphi_j(X); \sigma)$, $j = \overline{1,3}$, где $\varphi_j(X)$ — математическое ожидание, а σ — его среднее квадратическое отклонение.

При формировании выборки $V = (x^i, y^i,\ i = \overline{1,n})$ между случайными величинами X, Y использовались функциональные зависимости:

$$Y = \varphi_1(X) = 0{,}5X^2, \tag{8}$$

$$Y = \varphi_2(X) = 0{,}5X^3, \tag{9}$$

$$Y = \varphi_3(X) = 5\sin(X), \tag{10}$$

представленные на рис. 1.

Датчики случайных величин X и Y формировались на основе выражений

$$X = -3 + 6\varepsilon_1, \tag{11}$$

$$Y = \varphi_j(X) + \sigma\Big(\sum_{l=1}^{r} \varepsilon_2^l - 0{,}5r\Big)\frac{6}{\sqrt{3r}}, \qquad j = \overline{1,3}, \tag{12}$$

где ε_1, ε_2 — значения случайных величин с равномерной плотностью вероятности на интервале $[0; 1]$, значение параметра r принимается равным 12.

Иллюстрация полученных распределений при $n = 500$ для функциональных зависимостей (8), (9), (10) при различных значениях σ приведена на рис. 2.

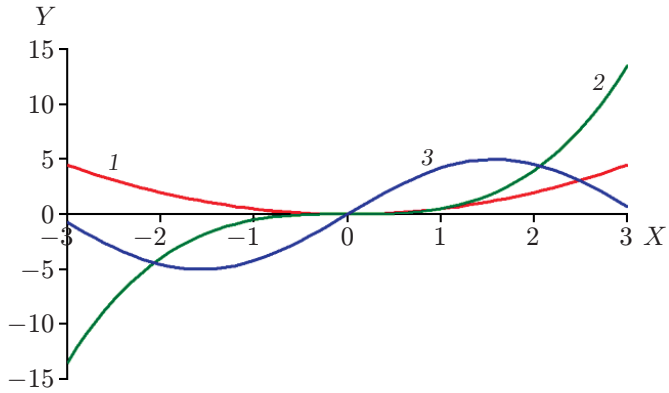

Рис. 1. Функциональные зависимости $\varphi_j(X)$, $j = \overline{1,3}$, между компонентами X, Y. Кривые *1, 2, 3* соответствуют выражениям (8)–(10)

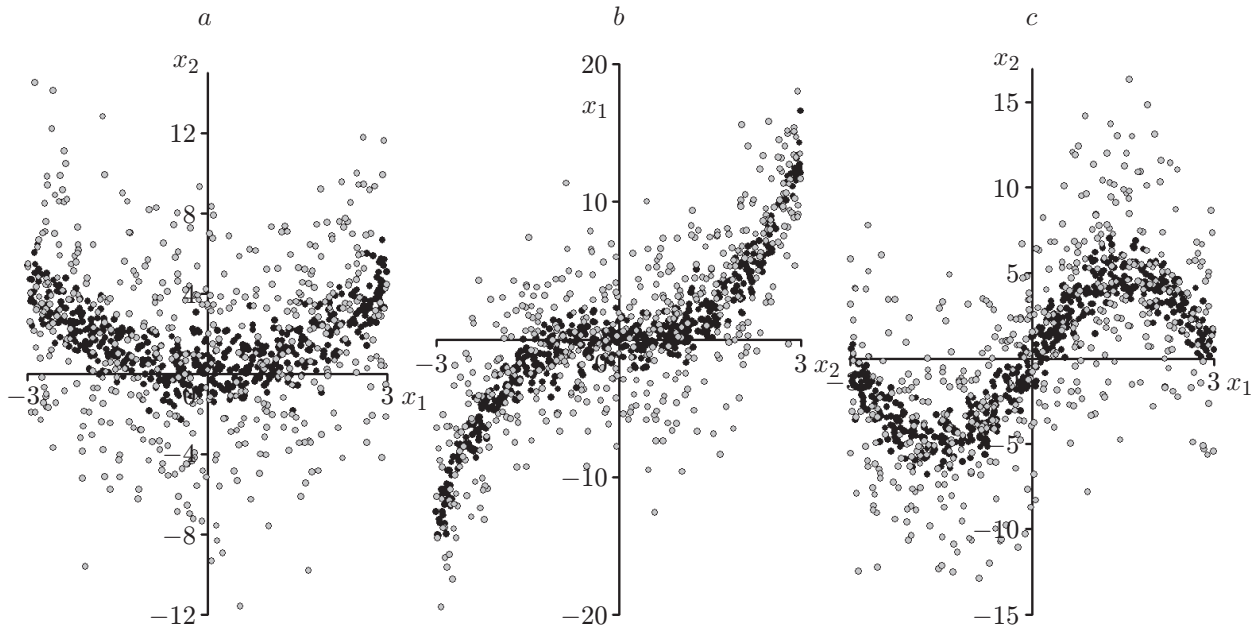

Рис. 2. Иллюстрация законов распределения исходных статистических данных V при $n = 500$ для функциональных зависимостей: *a* — (8), *b* — (9), *c* — (10). Чёрные точки соответствуют $\sigma = 1$, а серые — значению $\sigma = 4$

Объём n исходных статистических данных при организации вычислительных экспериментов определялся значениями 100, 200, 300, 400, 500. При конкретном объёме n исходных данных значения оценок вероятностей ошибок распознавания образов $\bar{\rho}_1$, $\bar{\rho}_2$ и D_{12} вычислялись 10 раз, а их результаты усреднялись и представлены в табл. 1–3.

Верхние элементы табл. 1 соответствуют значениям $\bar{\rho}_1$, а нижние — $\bar{\rho}_2$. Пороговые значения критерия Колмогорова — Смирнова для объёма исходных статистических данных $n = 100$, 200, 300, 400, 500 равны $D_\beta = 0{,}192$; 0,136; 0,111; 0,096; 0,086.

Проведём анализ результатов вычислительных экспериментов, которые соответствуют функциональной зависимости (8) (см. табл. 1). При малых значениях параметра σ в выражении (12) гипотеза о независимости случайных величин X, Y отвергается для всех значений $n \in [100, 500]$, что соответствует принятым условиям вычислительных экспериментов. Например, при $\sigma = 1$ и $n = 100$ значение $\bar{\rho}_1 = 0{,}84$, а $\bar{\rho}_2 = 0{,}159$. В этом случае

Таблица 1

Оценки вероятностей ошибок распознавания образов $\bar{\rho}_1$, $\bar{\rho}_2$ при функциональной зависимости (8) и различных значениях параметра σ

Объём выборки n	Значения среднего квадратического отклонения σ			
	1	2	3	4
100	0,84	0,624	0,524	0,471
	0,159	0,374	0,474	0,527
200	0,85	0,662	0,557	0,5235
	0,148	0,337	0,441	0,474
300	0,849	0,694	0,564	0,514
	0,15	0,305	0,435	0,485
400	0,855	0,688	0,582	0,526
	0,144	0,311	0,417	0,473
500	0,849	0,679	0,577	0,5442
	0,151	0,319	0,421	0,454

Таблица 2

Оценки вероятностей ошибок распознавания образов $\bar{\rho}_1$, $\bar{\rho}_2$ при функциональной зависимости (9) и различных значениях параметра σ

Объём выборки n	Значения среднего квадратического отклонения σ			
	1	2	3	4
100	0,956	0,904	0,848	0,786
	0,04	0,093	0,15	0,207
200	0,9595	0,9135	0,8605	0,8095
	0,0375	0,0845	0,138	0,19
300	0,9487	0,892	0,8637	0,8043
	0,0507	0,1063	0,1357	0,195
400	0,9485	0,8935	0,84175	0,812
	0,05	0,1045	0,1567	0,1873
500	0,9474	0,9042	0,8502	0,8178
	0,0518	0,095	0,1488	0,1804

отклонение $D_{12} = |\bar{\rho}_1 - \bar{\rho}_2| = 0{,}681$, а пороговое значение критерия Колмогорова — Смирнова $D_\beta = 0{,}192$ при риске $\beta = 0{,}05$ отвергнуть гипотезу $\bar{H}_0$: $\rho_1 = \rho_2$. Если $D_{12} > D_\beta$, то гипотеза $\bar{H}_0$ не выполняется. Из справедливости соотношения $\bar{\rho}_1 > \bar{\rho}_2$ следует условие зависимости случайных величин X, Y. При $\sigma = 1$ и $n = 500$ значение $\bar{\rho}_1 = 0{,}849$, а $\bar{\rho}_2 = 0{,}151$. В этом случае $D_{12} = 0{,}698$ и $D_\beta = 0{,}086$. Из соотношений $D_{12} > D_\beta$ и $\bar{\rho}_1 > \bar{\rho}_2$ следует вывод о зависимости случайных величин X и Y.

С ростом параметра σ наблюдается снижение значений $\bar{\rho}_1$ и увеличение $\bar{\rho}_2$, так как повышается влияние второй случайной составляющей с нормальным законом распределения $N(0;\sigma)$ выражения (12) на функциональную зависимость $\varphi_1(X)$ (8). Однако при $\sigma = 2$

Таблица 3

Оценки вероятностей ошибок распознавания образов $\bar{\rho}_1$, $\bar{\rho}_2$ при функциональной зависимости (10) и различных значениях параметра σ

Объём выборки n	Значения среднего квадратического отклонения σ			
	1	2	3	4
100	0,984	0,925	0,853	0,675
	0,016	0,074	0,145	0,325
200	0,9695	0,916	0,837	0,7545
	0,029	0,0835	0,1625	0,2435
300	0,96833	0,9	0,828	0,7647
	0,0317	0,0993	0,17	0,2343
400	0,96	0,903	0,836	0,7387
	0,04	0,097	0,1643	0,2605
500	0,9624	0,9056	0,8384	0,7592
	0,0376	0,0936	0,1614	0,2402

и $n \in [100, 500]$ значение $D_{12} > D_\beta$ и $\bar{\rho}_1 > \bar{\rho}_2$, что подтверждает зависимость случайных величин в соответствии с предложенной методикой.

При $\sigma = 4$ влияние второй составляющей выражения (12) на $\varphi_1(X)$ становится значимым и оказывает компенсирующее воздействие на рассматриваемую функциональную зависимость (8). Максимальные значения первой и второй составляющих выражения (12) сопоставимы и равны 4,5 и 4 соответственно. Поэтому в рассматриваемых условиях соотношения $D_{12} > D_\beta$ и $\bar{\rho}_1 > \bar{\rho}_2$, подтверждающие исходную зависимость, выполняются только при $n = 500$. Если $\sigma = 3$, то зависимость между случайными величинами подтверждается при $n \in [300, 500]$.

При усложнении функциональной зависимости (9), (10) между случайными величинами в условиях рассматриваемых вычислительных экспериментов предлагаемая методика

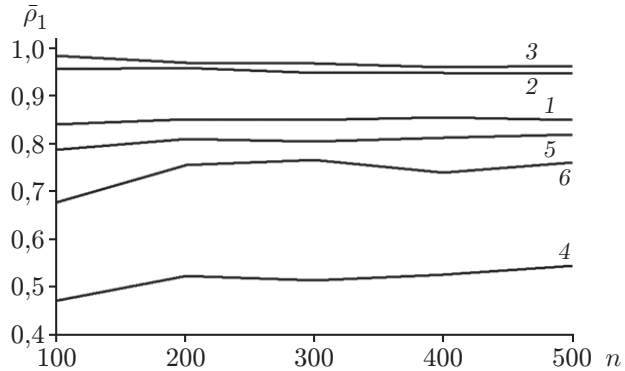

Рис. 3. Зависимость усреднённых оценок вероятностей ошибок $\bar{\rho}_1$ принадлежности элементов выборки V к условиям независимых случайных величин X, Y от объёма выборки n (кривые *1, 2, 3* соответствуют функциональным зависимостям (8), (9), (10) при значениях среднего квадратического отклонения $\sigma = 1$, а кривые *4, 5, 6* — значению $\sigma = 4$)

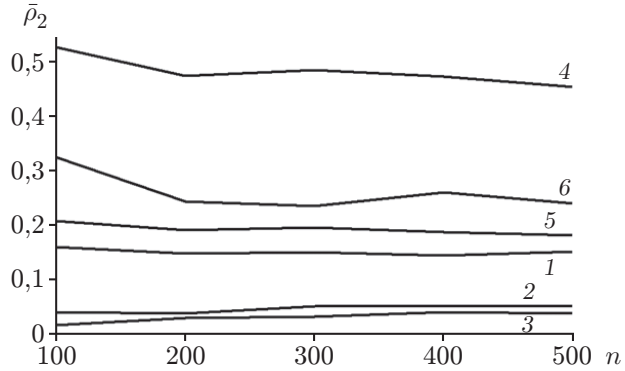

Рис. 4. Зависимость усреднённых оценок вероятностей ошибок $\bar{\rho}_2$ принадлежности элементов выборки V к условиям зависимых случайных величин X, Y от объёма выборки n. Кривые *1, 2, 3* соответствуют функциональным зависимостям (8), (9), (10) при значениях среднего квадратического отклонения $\sigma = 1$, а кривые *4, 5, 6* — значению $\sigma = 4$

достоверно отвергает гипотезу $\bar{H}_0$ о независимости X, Y, а оценки вероятностей ошибок распознавания образов $\bar{\rho}_2$ меньше значений $\bar{\rho}_1$. В соответствии с предлагаемой методикой случайные величины X, Y являются зависимыми, что согласуется с принятыми условиями вычислительных экспериментов.

Данный вывод подтверждается при анализе информации, представленной на рис. 3, 4.

Например, при параметре $\sigma = 1$, который определяет компенсирующее влияние на зависимость между X и Y, значения $\bar{\rho}_1$ увеличиваются при смене функциональных зависимостей от (8) до (10) (кривые *1, 2, 3*; см. рис. 3). В этих условиях значения $\bar{\rho}_2$ снижаются (кривые *1, 2, 3*; см. рис. 4), что подтверждает зависимость случайных величин X, Y. Предлагаемая методика является эффективной при различных объёмах исходных статистических данных. Этот вывод подтверждается малыми изменениями значений $\bar{\rho}_1$, $\bar{\rho}_2$ от объёма n исходных данных. При увеличении параметра σ до значения 4 отмеченные выше выводы сохраняются для функциональных зависимостей (9), (10) (кривые *5, 6*; см. рис. 3, 4). В этих условиях для функциональной зависимости (8) достоверное решение о зависимости случайных величин X, Y отмечается только при $n = 500$ (кривая *4*; см. рис. 3, 4).

Заключение. При проверке гипотезы о независимости случайных величин обоснована возможность использования непараметрического алгоритма распознавания образов, соответствующего критерию максимального правдоподобия. В отличие от традиционной постановки задачи классификации априори отсутствует обучающая выборка. Исходная информация представляется статистическими данными значений двухмерной случайной величины. Законы распределений случайных величин в классах оцениваются по исходным статистическим данным для условий их зависимости и независимости. При оптимальных коэффициентах размытости ядерных функций вычисляются оценки вероятностей ошибок распознавания образов в классах. Если они достоверно не отличаются, то исследуемые случайные величины являются независимыми. В противном случае принимается решение о зависимости случайных величин. Применение предлагаемого подхода позволяет обойти проблему декомпозиции области значений случайных величин на многомерные интервалы, что свойственно критерию Пирсона.

Эффективность предлагаемой методики оценивается по результатам вычислительных экспериментов. Исходная статистическая информация формировалась на основе нелинейных функциональных зависимостей между анализируемыми переменными и случайной

составляющей с нормальным законом распределения, среднее квадратическое отклонение σ которой изменялось. При квадратической функциональной зависимости между переменными и малых значениях параметра σ гипотеза о назависимости случайных величин отвергается для объёма статистических данных $n \in [100, 500]$. Когда значение параметра $\sigma \geqslant 3$, то компенсирующее влияние случайной составляющей оказывается соизмеримо со значениями квадратической функциональной зависимости. При $n \leqslant 200$ в принятых условиях вычислительного эксперимента предлагаемая методика достоверно не обнаруживает зависимости между случайными величинами. При усилении функциональной зависимости между переменными до кубической и синусоидальной эффективность предлагаемой методики подтверждается для $\sigma \in [1; 4]$ и $n \in [100, 500]$.

Перспективным направлением исследований является использование предлагаемой методики при формировании наборов зависимых и независимых случайных величин. Полученные результаты позволят с новых позиций решить проблему выбора наборов информативных признаков для решения задач принятия решений в условиях априорной неопределённости.

СПИСОК ЛИТЕРАТУРЫ

1. **Пугачев В. С.** Теория вероятностей и математическая статистика: Учеб. пособие. М: Физматлит, 2002. 496 с.
2. **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.
3. **Лапко А. В., Лапко В. А.** Сравнение эмпирической и теоретической функций распределения случайной величины на основе непараметрического классификатора // Автометрия. 2012. **48**, № 1. С. 45–49.
4. **Лапко А. В., Лапко В. А.** Методика проверки гипотез о распределениях многомерных спектральных данных с использованием непараметрического алгоритма распознавания образов // Компьютерная оптика. 2019. **43**, № 2. С. 238–244. DOI: 10.18287/2412-6179-2019-43-2-238-244.
5. **Лапко А. В., Лапко В. А.** Проверка гипотезы о независимости двухмерных случайных величин с использованием непараметрического алгоритма распознавания образов // Автометрия. 2021. **57**, № 2. С. 41–48. DOI: 10.15372/AUT20210205.
6. **Parzen E.** On estimation of a probability density function and mode // Ann. Math. Statist. 1962. **33**, N 3. P. 1065–1076.
7. **Епанечников В. А.** Непараметрическая оценка многомерной плотности вероятности // Теория вероятности и её применения. 1969. **14**, № 1. С. 156–161.
8. **Лапко А. В., Медведев А. В., Тишина Е. А.** К оптимизации непараметрических оценок // Сб. науч. тр. «Алгоритмы и программы для систем автоматизации экспериментальных исследований». Фрунзе: Илим, 1975. С. 105–116.
9. **Rudemo M.** Empirical choice of histogram and kernel density estimators // Scand. Journ. Statist. 1982. N 9. P. 65–78.
10. **Bowman A. W.** A comparative study of some kernel-based non-parametric density estimators // Journ. Statist. Comput. Simulation. 1982. **21**. P. 313–327.
11. **Hall P.** Large-sample optimality of least squares cross-validation in density estimation // Ann. Statist. 1983. **11**, N 4. P. 1156–1174.
12. **Jiang M., Provost S. B.** A hybrid bandwidth selection methodology for kernel density estimation // Journ. Statist. Comput. Simulation. 2014. **84**, N 3. P. 614–627. DOI: 10.1080/00949655.2012.721366.

13. **Dutta S.** Cross-validation revisited // Commun. Statistics — Simulation and Comput. 2016. **45**, N 2. P. 472–490. DOI: 10.1080/03610918.2013.862275.
14. **Heidenreich N.-B., Schindler A., Sperlich S.** Bandwidth selection for kernel density estimation: A review of fully automatic selectors // AStA Adv. Statist. Anal. 2013. **97**, N 4. P. 403–433. DOI: 10.1007/s10182-013-0216-y.
15. **Li Q., Racine J. S.** Nonparametric Econometrics: Theory and Practice. Princeton: Princeton University Press, 2007. 768 p.
16. **Duin R. P. W.** On the choice of smoothing parameters for Parzen estimators of probability density functions // IEEE Trans. Comp. 1976. **25**, Iss. 11. P. 1175–1179.
17. **Botev Z. I., Kroese D. P.** Non-asymptotic bandwidth selection for density estimation of discrete data // Methodology and Computing in Appl. Probability. 2008. **10**, N 3. P. 435–451.
18. **Лапко А. В., Лапко В. А.** Методика быстрого выбора коэффициентов размытости в непараметрическом классификаторе, соответствующем критерию максимума апостериорной вероятности // Автометрия. 2019. **55**, № 6. С. 76–86. DOI: 10.15372/AUT20190610.
19. **Лапко А. В., Лапко В. А.** Модифицированный алгоритм быстрого определения коэффициента размытости ядерной оценки плотности вероятности // Автометрия. 2020. **56**, № 6. С. 11–18. DOI: 10.15372/AUT20200602.
20. **Scott D. W.** Multivariate Density Estimation: Theory, Practice, and Visualization. New Jersey: John Wiley & Sons, 2015. 384 p.
21. **Sheather S. J.** Density Estimation // Statist. Sci. 2004. **19**, N 4. P. 588–597. DOI: 10.1214/088342304000000297.
22. **Silverman B. W.** Density Estimation for Statistics and Data Analysis. London: Chapman and Hall, 1986. 175 p.
23. **Лапко А. В., Лапко В. А.** Оценивание нелинейного функционала от плотности вероятности при оптимизации непараметрических решающих функций // Измерительная техника. 2021. № 1. С. 14–20. DOI: 10.32446/0368-1025it.2021-1-14-20.
24. **Шаракшанэ А. С., Железнов И. Г., Ивницкий В. А.** Сложные системы. М.: Высш. шк., 1977. 248 с.

Поступила в редакцию 31.05.2021
После доработки 30.08.2021
Принята к публикации 03.09.2021