

Сравнение методики проверки гипотезы о независимости двумерных случайных величин, основанной на непараметрическом классификаторе

А.В. Лапко ^{I,II}, В.А. Лапко ^{I,II}, А.В. Бахтина ^{II}

^I Институт вычислительного моделирования СО РАН, г. Красноярск, Россия

^{II} Сибирский государственный университет науки и технологий им. акад. М.Ф. Решетнева, г. Красноярск, Россия

Аннотация. Рассматриваются свойства новой 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 анализируемых 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. Результаты применения предлагаемой методики сравниваются с критерием Пирсона и коэффициентами корреляции Пирсона, Спирмена и Кендалла. При реализации критерия Пирсона используется формула оптимальной дискретизации области значений двумерной случайной величины. Их эффективность при усложнении зависимости между случайными величинами и изменении объема исходных статистических данных исследуется методом вычислительного эксперимента.

Ключевые слова: проверка гипотезы о независимости случайных величин, двумерные случайные величины, непараметрический алгоритм распознавания образов, ядерная оценка плотности вероятности, критерий Пирсона, зависимые случайные величины, коэффициенты корреляции Пирсона, Спирмена и Кендалла.

DOI 10.14357/20718594220105

Введение

Независимость либо зависимость случайных величин является необходимой информацией в задаче выбора значимых переменных при синтезе 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 задача разработки новой методики проверки гипотезы о независимости случайных величин,

которая обеспечивает обход проблемы декомпозиции области их значений на многомерные интервалы, является актуальной.

Решение подобной задачи при проверке гипотезы о тождественности законов распределения случайных величин на основе использования непараметрического алгоритма распознавания образов, предложено в работах [2-4]. Показана возможность ее замены на задачу проверки гипотезы о равенстве ошибки распознавания образов определенному пороговому значению. Обучающая выборка при синтезе непараметрического алгоритма распознавания образов формируется по статистическим данным, характеризующим законы распределения

✉ Лапко Александр Васильевич. E-mail: lapko@icm.krasn.ru

сравниваемых случайных величин. Такой 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.

Цель статьи состоит в сравнении эффективности предлагаемой методики проверки гипотезы о независимости случайных величин с традиционным критерием Пирсона и коэффициентами корреляции Пирсона, Спирмена и Кендалла при усложнении функциональных зависимостей между случайными величинами и изменении объема статистических данных.

1. Непараметрический алгоритм распознавания образов, соответствующий критерию максимального правдоподобия

Пусть имеется выборка $V = (x^i, i = \overline{1, n})$ объема n , составленная из наблюдений двумерной случайной величины $x = (x_1, x_2)$. Выборка V извлекается из генеральных совокупностей, характеризующихся априори неизвестными плотностями вероятностей $p(x_1)p(x_2)$ или $p(x_1, x_2)$. Необходимо по статистическим данным V проверить гипотезу

$$H_0: p(x_1, x_2) \equiv p(x_1)p(x_2) \quad (1)$$

о независимости случайных величин x_1, x_2 .

Для проверки гипотезы H_0 будем решать двухальтернативную задачу распознавания образов. Под классами Ω_1 и Ω_2 понимаются области определения плотностей вероятностей $p(x_1)p(x_2)$ и $p(x_1, x_2)$.

В отличие от традиционной постановки задачи распознавания образов при синтезе решающего правила классификации 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 $p(x_1, x_2)$ и $p(x_1)p(x_2)$, используя их непараметрические оценки типа Розенблатта – Парзена [6, 7]:

$$\bar{p}(x_1, x_2) = \frac{1}{n c_1 c_2} \sum_{i=1}^n \Phi\left(\frac{x_1 - x_1^i}{c_1}\right) \Phi\left(\frac{x_2 - x_2^i}{c_2}\right), \quad (2)$$

$$\bar{p}(x_1) \bar{p}(x_2) = \frac{1}{n^2 c_1 c_2} \sum_{i=1}^n \sum_{j=1}^n \Phi\left(\frac{x_1 - x_1^i}{c_1}\right) \Phi\left(\frac{x_2 - x_2^j}{c_2}\right), \quad (3)$$

где c_v – коэффициенты размытости ядерных функций $\Phi(u_v)$, $v=1, 2$.

В статистиках (2), (3) ядерные функции $\Phi(u_v)$ удовлетворяют условиям:

$$\Phi(u_v) = \Phi(-u_v), \quad 0 \leq \Phi(u_v) < \infty, \quad \int \Phi(u_v) du_v = 1, \\ \int u^m \Phi(u_v) du_v < \infty, \quad 0 \leq m < \infty, \quad v=1, 2.$$

Здесь и далее бесконечные пределы интегрирования опускаются. Тогда непараметрическое решающее правило классификации случайных величин $x = (x_1, x_2)$, соответствующее критерию максимального правдоподобия, запишется как:

$$\bar{m}(x): \begin{cases} x \in \Omega_1, & \text{если } \bar{p}(x_1, x_2) < \bar{p}(x_1) \bar{p}(x_2) \\ x \in \Omega_2, & \text{если } \bar{p}(x_1, x_2) > \bar{p}(x_1) \bar{p}(x_2). \end{cases} \quad (4)$$

Оптимальные коэффициенты размытости c_v ядерных функций непараметрических оценок плотностей вероятностей $\bar{p}(x_v)$ будем определять из условия минимума асимптотических выражений критериев

$$W(c_v) = \int (\bar{p}(x_v) - p(x_v))^2 dx_v, \quad (5)$$

которые характеризует меру близости между $\bar{p}(x_v)$ и $p(x_v)$, $v=1, 2$.

В работах [8-15] обоснована возможность определения оптимальных значений c_v путем минимизации выражений:

$$\bar{W}(c_v) = \frac{1}{n^2 c_v^2} \sum_{j=1}^n \sum_{i=1}^n \int \Phi\left(\frac{x_v - x_v^j}{c_v}\right) \Phi\left(\frac{x_v - x_v^i}{c_v}\right) dx_v - \\ - \frac{2}{n^2 c_v} \sum_{j=1}^n \sum_{i=1, i \neq j}^n \Phi\left(\frac{x_v^j - x_v^i}{c_v}\right), \quad v=1, 2. \quad (6)$$

По аналогии с выражением (6) нетрудно определить критерий выбора оптимальных коэффициентов размытости непараметрической статистики $\bar{p}(x_1, x_2)$ (2).

Оптимизацию непараметрического решающего правила (4) по коэффициентам размытости ядерных функций можно упростить, если положить в статистиках (2), (3) значения $c_v = c\bar{\sigma}_v$, $v=1, 2$. Здесь $\bar{\sigma}_v$ – оценка среднего квадратического отклонения случайной величины x_v , вычисляемая по выборке V . Данное утверждение является очевидным, так как большей длине интервала значений x_v соответствует больший коэффициент размытости c_v ядерных функций $\Phi(u_v)$, $v=1, 2$. Поэтому появляется возможность оптимизацию непараметрического алгоритма распознавания образов (4) проводить лишь по одному параметру с коэффициентов размытости ядерных функций. Подобный подход использовался при построении быстрых процедур оптимизации непараметрических оценок плотностей вероятностей ядерного типа [16-21].

2. Методика проверки гипотезы о независимости компонент двумерной случайной величины

Непараметрический алгоритм распознавания образов (4) основан на проверке соотношений между ядерными оценками плотностей вероятностей $\bar{p}(x_1)\bar{p}(x_2)$ и $\bar{p}(x_1, x_2)$. Для ситуаций первого класса Ω_1 , в которых соотношение $\bar{p}(x_1)\bar{p}(x_2) > \bar{p}(x_1, x_2)$, подтверждает справедливость предположения о независимости случайных величин x_1 и x_2 . В области определения непараметрической оценки плотности вероятности $\bar{p}(x_1, x_2)$ при выполнении соотношения $\bar{p}(x_1)\bar{p}(x_2) < \bar{p}(x_1, x_2)$ следует зависимость случайных величин. Выполнение гипотезы (1) определяет границу в области значений случайных величин x_1 и x_2 , разделяющих предположения о независимости либо зависимости x_1, x_2 . Поэтому методика проверки гипотезы о независимости случайных величин предполагает выполнение следующих действий:

1. В соответствии с рекомендациями предыдущего раздела осуществить синтез непараметрического алгоритма распознавания образов (4).

2. Определить оценки вероятностей ошибок распознавания образов $\bar{p}_1, \bar{p}_2$ решающего правила (4) по исходным статистическим данным

V при оптимальных коэффициентах размытости ядерных статистик $\bar{p}(x_1)\bar{p}(x_2), \bar{p}(x_1, x_2)$.

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

$$\bar{p}_t = \frac{1}{n} \sum_{j=1}^n 1(\delta(j), \bar{\delta}(j)), t=1, 2, \quad (7)$$

где $\delta(j)$ – указания типа $x^j = (x_1^j, x_2^j) \in \Omega_t$, а $\bar{\delta}(j)$ – «решение» алгоритма (4) о принадлежности ситуации x^j к одному из классов Ω_t , $t=1, 2$.

При вычислении $\bar{p}_t$ в соответствии с методикой «скользящего экзамена» ситуация $x^j = (x_1^j, x_2^j)$ из выборки V , которая подается на контроль в алгоритм (4), исключается из процесса формирования статистик (2), (3).

Индикаторная функция в формуле (7) 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 $\bar{p}_1$ и $\bar{p}_2$ в предположении, что элементы выборки V принадлежат классам Ω_1, Ω_2 соответственно. Тогда гипотеза H_0 справедлива, если $\bar{p}_1 < \bar{p}_2$. В противном случае при $\bar{p}_2 < \bar{p}_1$ случайные величины x_1 и x_2 являются зависимыми.

При ограниченных объемах n выборки V возникает задача доверительного оценивания вероятностей ошибок распознавания образов p_1 и p_2 . Для ее решения используется традиционная методика проверки гипотезы о равенстве двух вероятностей p_1, p_2 [1]. В качестве наблюдаемого значения критерия проверки гипотезы используется статистика:

$$U = \frac{|\bar{p}_1 - \bar{p}_2|}{\sqrt{\frac{\bar{p}_1 + \bar{p}_2}{2} \left(1 - \frac{\bar{p}_1 + \bar{p}_2}{2}\right) \frac{2}{n}}}.$$

При уровне значимости 0,05 проверяемая гипотеза выполняется, если $U < 1,96$.

3. Анализ результатов вычислительных экспериментов

Сравним эффективность предлагаемой методики с традиционным критерием Пирсона

и коэффициентами корреляции Пирсона, Спирмена и Кендалла при различных условиях зависимости между случайными величинами x_1 , x_2 и изменении объема n статистических данных. Значения x_1 формируются с помощью датчика с равномерным законом распределения $p(x_1)=R(0; 1,732)$ с математическим ожиданием 0 и средним квадратическим отклонением 1,732.

Значения случайной величины x_2 определяются нормальным законом распределения $N(\varphi_j(x_1), \sigma)$, $j = \overline{1, 4}$, где $\varphi_j(x_1)$ - математическое ожидание x_2 , а σ - среднее квадратическое отклонение x_2 .

В случае независимых случайных величин x_1 , x_2 выборка $V = (x_1^i, x_2^i, i = \overline{1, n})$ формируется с использованием равномерного $p(x_1)=R(0;1,732)$ и нормального $p(x_2)=N(0,\sigma)$ законами распределения.

При формировании выборки V для зависимых случайных величин x_1 и x_2 использовались функциональные зависимости:

$$\varphi_1(x_1) = x_1 \quad (8)$$

$$\varphi_2(x_1) = 0,5x_1^2, \quad (9)$$

$$\varphi_3(x_1) = 0,5x_1^3, \quad (10)$$

$$\varphi_4(x_1) = 3\sin(x_1). \quad (11)$$

Датчики зависимых случайных величин x_1 и x_2 формировались на основе выражений:

$$x_1 = -3 + 6\varepsilon_1,$$

$$x_2 = \varphi_j(x_1) + \sigma \left(\sum_{l=1}^r \varepsilon_l^j - 0,5r \right) 6/\sqrt{3r}, \quad j = \overline{1, 6},$$

где $\varepsilon_1, \varepsilon_2$ - значения случайных величин с равномерной плотностью вероятности на интервале $[0; 1]$. Значение параметра r принимается равным 12.

Графическая иллюстрация полученных распределений при $n=500$ для независимых и зависимых случайных величин (8) - (11) для различных значений σ представлены на Рис. 1.

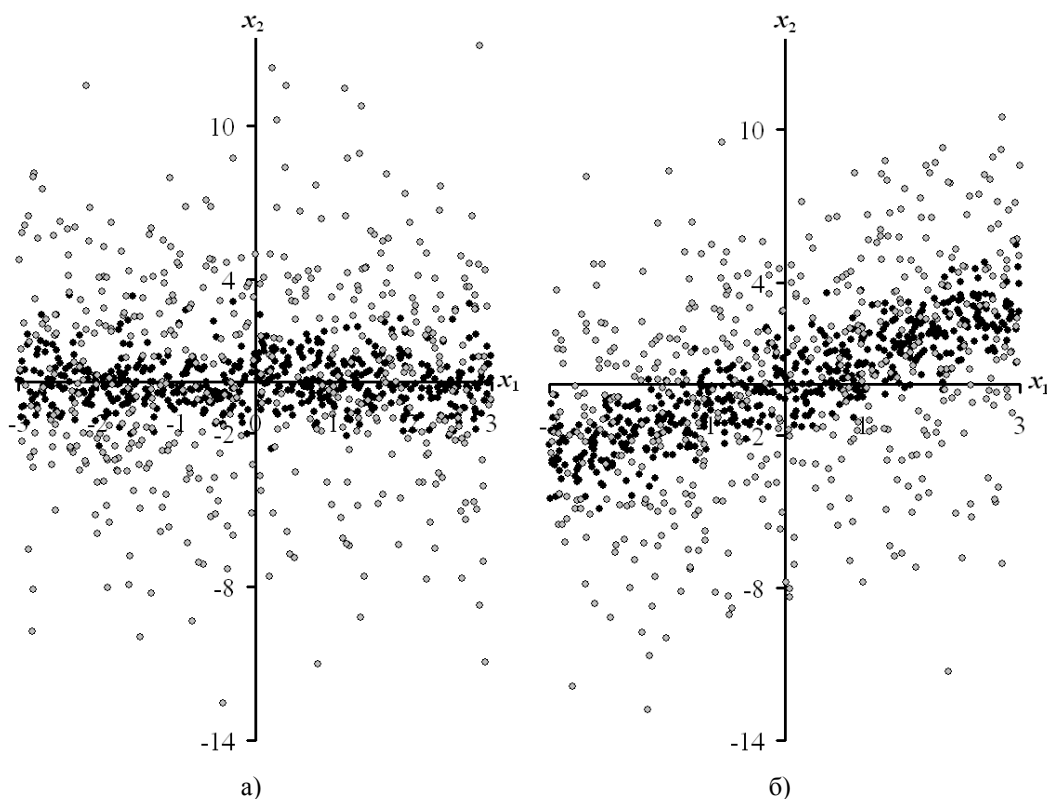


Рис. 1 (начало). Иллюстрация законов распределения статистических данных V при $n=500$ для независимых а) и зависимых случайных величин x_1 и x_2 . Фрагмент б) соответствует зависимости (8), в) – (9), з) – (10), д) – (11). Темные точки соответствуют значению $\sigma=1$, а серые – $\sigma=4$

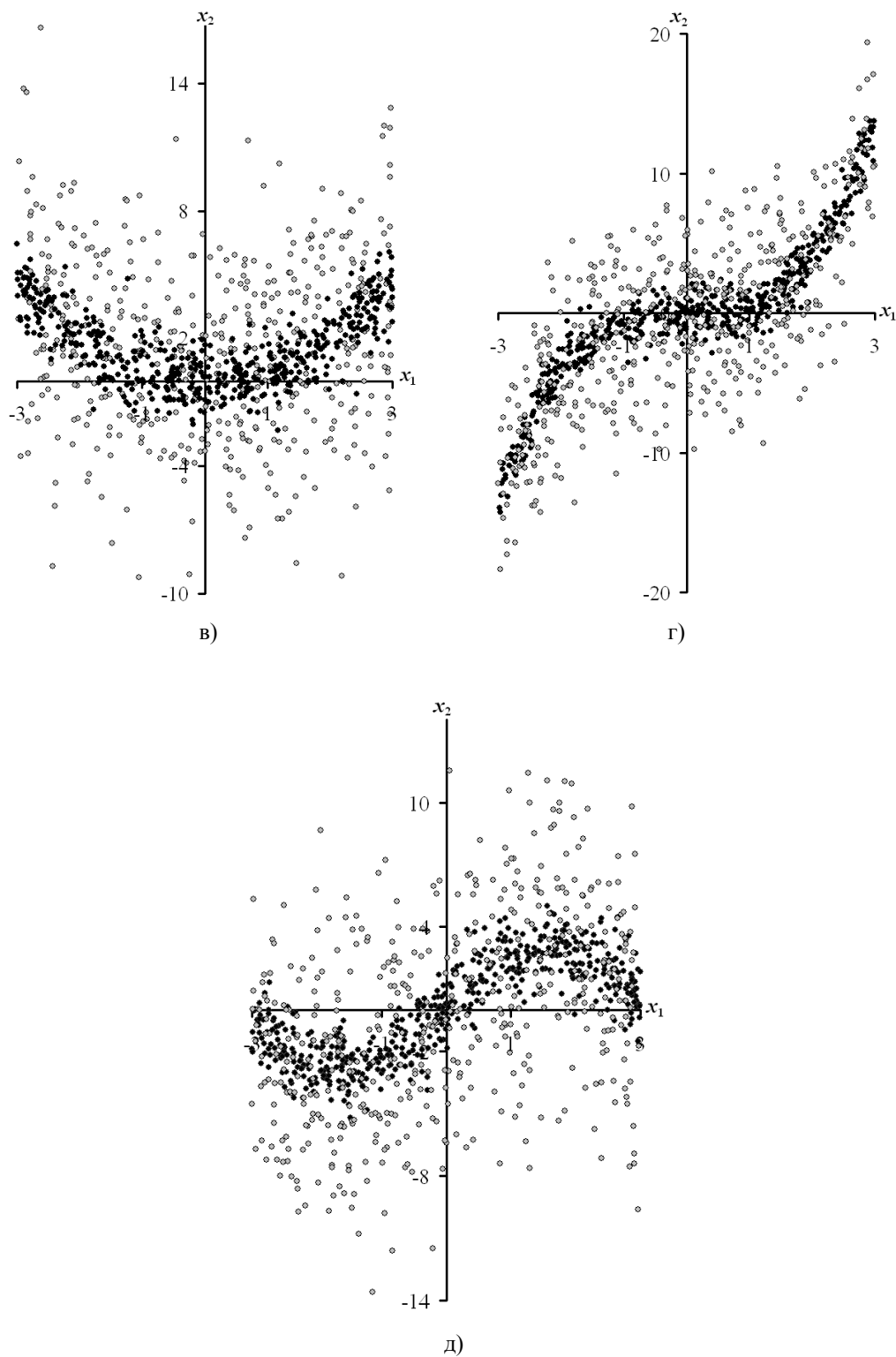


Рис. 1 (конец). Иллюстрация законов распределения статистических данных V при $n=500$ для независимых a) и зависимых случайных величин x_1 и x_2 . Фрагмент b) соответствует зависимости (8), g) – (9), z) – (10), d) – (11). Темные точки соответствуют значению $\sigma=1$, а серые – $\sigma=4$

Объем n статистических данных при организации вычислительных экспериментов определялся значениями 100, 300, 500. При конкретном объеме n исходных данных значения

оценок вероятностей ошибок распознавания образов $\bar{\rho}_1$, $\bar{\rho}_2$ вычисляются 10 раз, а их результаты усреднялись. Полученные результаты представлены в Табл. 1–5.

Табл. 1. Результаты вычислительных экспериментов при анализе независимых случайных величин x_1 и x_2 с использованием предлагаемой методики и критерия Пирсона

Объем выборки n	Значение σ	Значения		Значение критерия U	Количество интервалов $\bar{N}$	Значение χ^2	Порог χ^2
		$\bar{\rho}_1$	$\bar{\rho}_2$				
100	1	0,405	0,594	2,673	9	3,332	9,488
	2	0,434	0,561	1,796	9	3,849	9,488
	3	0,46	0,537	1,089	9	3,782	9,488
	4	0,416	0,582	2,348	9	5,064	9,488
300	1	0,389	0,61	5,413	16	7,38	16,919
	2	0,397	0,602	5,021	16	6,317	16,919
	3	0,401	0,597	4,801	16	9,618	16,919
	4	0,393	0,606	5,217	16	8,278	16,919
500	1	0,399	0,6	6,356	25	16,801	26,296
	2	0,404	0,595	6,040	25	15,589	26,296
	3	0,409	0,591	5,755	25	17,033	26,296
	4	0,443	0,557	3,605	25	19,063	26,296

Табл. 2. Результаты вычислительных экспериментов при анализе зависимых случайных величин x_1 и x_2 в условиях (8) с использованием предлагаемой методики и критерия Пирсона

Объем выборки n	Значение σ	Значения		Значение критерия U	Количество интервалов $\bar{N}$	Значение χ^2	Порог χ^2
		$\bar{\rho}_1$	$\bar{\rho}_2$				
100	1	0,89	0,11	11,031	16	98,271	16,919
	2	0,678	0,32	5,063	9	37,001	9,488
	3	0,641	0,359	3,988	9	21,59	9,488
	4	0,524	0,475	0,693	9	11,418	9,488
300	1	0,893	0,107	19,253	25	320,832	26,296
	2	0,759	0,24	12,713	25	150,073	26,296
	3	0,661	0,338	7,912	16	77,566	16,919
	4	0,6	0,399	4,923	16	53,13	16,919
500	1	0,883	0,116	24,255	36	542,669	37,652
	2	0,76	0,239	16,475	25	225,628	26,296
	3	0,654	0,346	9,740	25	129,297	26,296
	4	0,587	0,413	5,502	25	97,148	26,296

Табл. 3. Результаты вычислительных экспериментов при анализе зависимых случайных величин x_1 и x_2 в условиях (9) с использованием предлагаемой методики и критерия Пирсона

Объем выборки n	Значение σ	Значения		Значение критерия U	Количество интервалов $\bar{N}$	Значение χ^2	Порог χ^2
		$\bar{\rho}_1$	$\bar{\rho}_2$				
100	1	0,802	0,195	8,584	16	36,935	16,919
	2	0,656	0,34	4,469	9	16,818	9,488
	3	0,577	0,422	2,192	9	11,242	9,488
	4	0,523	0,474	0,693	9	8,479	9,488
300	1	0,853	0,147	17,293	25	182,95	26,296
	2	0,695	0,304	9,578	16	79,711	16,919
	3	0,574	0,425	3,650	16	42,513	16,919
	4	0,503	0,496	0,171	16	25,046	16,919
500	1	0,85	0,15	22,136	25	296,97	26,296
	2	0,672	0,328	10,878	25	143,027	26,296
	3	0,56	0,44	3,795	25	67,108	26,296
	4	0,518	0,482	1,138	25	46,181	26,296

Табл. 4. Результаты вычислительных экспериментов при анализе зависимых случайных величин x_1 и x_2 в условиях (10) с использованием предлагаемой методики и критерия Пирсона

Объем выборки n	Значение σ	Значения		Значение критерия U	Количество интервалов $\bar{N}$	Значение χ^2	Порог χ^2
		$\bar{\rho}_1$	$\bar{\rho}_2$				
100	1	0,958	0,037	13,025	25	149,043	26,296
	2	0,887	0,105	11,060	16	86,348	16,919
	3	0,846	0,147	9,886	16	75,131	16,919
	4	0,786	0,213	8,103	9	60,192	9,488
300	1	0,943	0,054	21,776	49	581,97	50,998
	2	0,898	0,102	19,498	25	404,263	26,296
	3	0,869	0,129	18,126	25	274,097	26,296
	4	0,804	0,195	14,917	25	217,606	26,296
500	1	0,952	0,048	28,587	49	1178,207	50,998
	2	0,907	0,092	25,773	36	742,984	37,652
	3	0,853	0,146	22,357	36	564,252	37,652
	4	0,811	0,188	19,701	25	387,358	26,296

Табл. 5 (начало). Результаты вычислительных экспериментов при анализе зависимых случайных величин x_1 и x_2 в условиях (11) с использованием предлагаемой методики и критерия Пирсона

Объем выборки n	Значение σ	Значения		Значение критерия U	Количество интервалов $\bar{N}$	Значение χ^2	Порог χ^2
		$\bar{\rho}_1$	$\bar{\rho}_2$				
100	1	0,932	0,068	12,219	16	75,386	16,919
	2	0,78	0,22	7,920	9	43,148	9,488
	3	0,657	0,34	4,483	9	24,937	9,488
	4	0,605	0,395	2,970	9	14,953	9,488

Табл. 5 (продолжение). Результаты вычислительных экспериментов при анализе зависимых случайных величин x_1 и x_2 в условиях (11) с использованием предлагаемой методики и критерия Пирсона

Объем выборки n	Значение σ	Значения		Значение критерия U	Количество интервалов $\bar{N}$	Значение χ^2	Порог χ^2
		$\bar{p}_1$	$\bar{p}_2$				
300	1	0,927	0,073	20,919	25	287,32	26,296
	2	0,807	0,192	15,064	25	168,061	26,296
	3	0,674	0,326	8,524	16	86,826	16,919
	4	0,606	0,394	5,193	16	60,915	16,919
500	1	0,927	0,072	27,037	36	541,354	37,652
	2	0,802	0,197	19,132	25	262,034	26,296
	3	0,699	0,3	12,617	25	172,528	26,296
	4	0,63	0,37	8,222	25	116,646	26,296

При проверке гипотезы о независимости компонент двухмерной случайной величины на основе критерия Пирсона используются результаты оптимального выбора количества интервалов дискретизации [22, 23]:

$$N^* = \left(\frac{3}{4} \Delta_1 \Delta_2 \|p(x_1, x_2)\|^2 n \right)^{1/2}, \quad (12)$$

где значение $\|p(x_1, x_2)\|^2 = \iint p^2(x_1, x_2) dx_1 dx_2$, а Δ_v – интервал изменения случайной величины x_v , $v=1, 2$.

Результат (12) получен путем исследования асимптотических свойств регрессионной оценки плотности вероятности, синтез которой основывается на дискретизации области ее определения для обхода проблемы обработки больших объемов статистических данных. Оценивание функционала от плотности вероятности $p(x_1, x_2)$ осуществляется по выборке V с использованием статистики:

$$\begin{aligned} \|p(x_1, x_2)\|^2 &\approx \iint \bar{p}(x_1, x_2) p(x_1, x_2) dx_1 dx_2 = \\ &= \frac{1}{n^2 c_1 c_2} \sum_{j=1}^n \sum_{\substack{i=1 \\ i \neq j}}^n \Phi\left(\frac{x_1^j - x_1^i}{c_1}\right) \Phi\left(\frac{x_2^j - x_2^i}{c_2}\right). \end{aligned} \quad (13)$$

Условие $i \neq j$ в выражении (13) необходимо для обеспечения свойства несмещенности оценки математического ожидания непараметрической оценки плотности вероятности $\bar{p}(x_1, x_2)$.

Реализация теоретического результата (12) предполагает выполнение следующих действий. Вычислить значение $\bar{N} = (N^*)^{1/2}$. Определить количество интервалов дискретизации $\bar{N}_v$

каждой компоненты x_v , $v=1, 2$ случайной величины $x=(x_1, x_2)$. Здесь $\bar{N}_v$ соответствует ближайшему к $\bar{N}$ целому числу. Общее количество интервалов дискретизации области значений случайной величины x соответствует значению $\bar{N} = \bar{N}_1 \bar{N}_2$.

Для анализа результатов вычислительных экспериментов введем обозначения. Пусть R_1 определяет отношение χ^2 критерия и его порогового значения, а R_2 соответствует отношению критерия U и порога 1,96 при уровне значимости 0,05.

При анализе данных Табл. 1 предлагаемая методика и критерий Пирсона в основном обнаруживают свойство независимости случайных величин x_1 и x_2 для значений $\sigma \in [1, 4]$ и $n \in [100, 500]$. При использовании критерия Пирсона отношение $R_1 < 1$ для $\sigma \in [1, 4]$ и $n \in [100, 500]$, что подтверждает независимость случайных величин x_1 и x_2 . Например, при $n=100$ значение $R_1 \in [0,35; 0,53]$, а в условиях $n=500$ отношение $R_1 \in [0,63; 0,72]$. Предлагаемая методика проверки гипотезы о независимости случайных величин допускает ошибки при объеме выборки $n=100$ и $\sigma=2, \sigma=3$, для которых $R_2=0,91$ и $R_2=0,55$ соответственно. При увеличении объема n статистических данных рассматриваемая методика безошибочно подтверждает гипотезу о независимости x_1 и x_2 . Например, при $n=500$ значение $R_2 \in [1,84; 3,24]$. Наблюдаемые ошибки при проверке рассматриваемой гипотезы на основе предлагаемой методики объясняются оцениванием плотностей вероятностей в условиях малого объема статисти-

стических данных. При использовании критерия Пирсона оцениваются вероятности принадлежности значений случайных величин двумерным интервалам. Поэтому при удачном выборе количества интервалов дискретизации критерий Пирсона при малых объемах статистических данных имеет преимущество. Значения коэффициентов корреляции Пирсона, Спирмена, Кендалла при $\sigma \in [1, 4]$ и $n \in [100, 500]$ принадлежат интервалам $[-0,061; 0,03]$, $[-0,063; 0,036]$, $[-0,043; 0,023]$ соответственно.

В условиях (8) критерий Пирсона на основе используемой процедуры дискретизации области значений случайных величин безошибочно определяет зависимость между x_1 и x_2 (Табл. 2). Например, при $n=100$ и $\sigma \in [1, 4]$ значения $R_1 \in [1,2; 5,8]$. Если $n=500$, то $R_1 \in [3,69; 14,4]$. Причем, большие значения R_1 соответствуют малым значениям σ . В данных условиях линейной зависимости между x_1 и x_2 предлагаемая методика допускает ошибочное решение только при $n=100$ и $\sigma=4$, когда $R_2=0,35$. При $n=500$ значения $R_2 \in [2,8; 12,37]$. При этом коэффициенты корреляции Пирсона, Спирмена, Кендалла принимают значения в интервалах $[0,377; 0,873]$, $[0,363; 0,88]$, $[0,248; 0,688]$.

При квадратической зависимости между x_1 и x_2 в условиях (9) результаты применения предлагаемой методики и критерия Пирсона сопоставимы при $\sigma \in [1, 3]$ и $n \in [100, 500]$ (Табл. 3). Например, при $n=100$ и $\sigma \in [1, 3]$ значения $R_1 \in [1,18; 2,18]$, а $R_2 \in [1,12; 4,38]$. Если $n=500$, то значения $R_1 \in [1,75; 11,29]$, а $R_2 \in [1,93; 11,3]$. Предлагаемая методика принимает ошибочное решение при $\sigma=4$ и $n \in [100, 500]$. Данный факт можно объяснить особенностями квадратической зависимости и большим значениям $\sigma=4$. В этих условиях плотности вероятностей $p(x_1)p(x_2)$ и $p(x_1, x_2)$ трудноразличимы (Рис. 1, в). В этих условиях при $n=100$, $\sigma=4$ и критерий Пирсона не обнаруживает зависимость между x_1 и x_2 . Коэффициенты корреляции Пирсона, Спирмена, Кендалла принимают значения в интервалах $[-0,066; 0,066]$, $[-0,052; 0,045]$, $[-0,034; 0,022]$.

Для условий вычислительных экспериментов, определяемых выражением (10), результаты предлагаемой методики и критерия Пирсона сопоставимы и принимают безошибочное решение (Табл. 4). Например, при $n=100$ и $\sigma \in [1, 4]$ значения $R_1 \in [5,67; 6,34]$, а $R_2 \in [4,13; 6,65]$. При $n=500$

$R_1 \in [14,73; 23,1]$, а $R_2 \in [10,05; 14,57]$. Зависимость между x_1 и x_2 частично обнаруживается с использованием коэффициентов корреляции Пирсона, Спирмена, Кендалла, значения которых принадлежат интервалам $[0,706; 0,907]$, $[0,679; 0,927]$, $[0,505; 0,792]$.

Если зависимость между случайными величинами x_1 и x_2 характеризуется выражением (11), то предлагаемая методика и критерий Пирсона обнаруживает эту зависимость при $\sigma \in [1, 4]$ и $n \in [100, 500]$ (Табл. 5). Например, отношение χ^2 - критерия и порогового его значения при $n=100$ и $\sigma=1$ равно значению 4,15, а при $\sigma=4$ это отношение определяется значением 1,57. В этих условиях предлагаемая методика при сравнении ошибок распознавания образов $\bar{p}_1$, $\bar{p}_2$ отношение наблюдаемого значения U и порога 1,96 принадлежит интервалу $[1,51; 6,23]$. Данное отношение для $n=500$ принимает значения в интервале $[4,19; 13,79]$. Значения коэффициентов корреляции Пирсона, Спирмена, Кендалла при $\sigma=1$ и $n \in [100, 500]$ изменяется в пределах $[0,739; 0,756]$, $[0,734; 0,75]$, $[0,498; 0,522]$.

Заключение

Методика проверки гипотезы о независимости двумерной случайной величины основывается на использовании непараметрического алгоритма распознавания образов, соответствующего критерию максимального правдоподобия. Ее использование позволяет обойти проблему дискретизации области значений случайных величин, которая свойственна критерию Пирсона. По результатам вычислительных экспериментов предлагаемая методика и критерий Пирсона сопоставимы. При малых объемах n статистических данных и больших значениях σ наблюдается преимущество критерия Пирсона для линейной и квадратической зависимости между случайными величинами, если используется выбор оптимального количества интервалов дискретизации (12). Данный факт объясняется различием условий сходимости непараметрических оценок плотностей вероятностей и частот принадлежности случайных величин двумерным интервалам. Коэффициенты корреляции Пирсона, Спирмена, Кендалла позволяют обнаружить линейную

зависимость между случайными величинами. Линейная зависимость обнаруживается коэффициентами корреляции Пирсона, Спирмена в основном при малых значениях σ . По эффективности им уступает коэффициент корреляции Кендалла.

Перспективное направление исследований состоит в использовании предлагаемой методики при анализе переменных неоднозначных зависимостей в условиях воздействия на них случайных помех.

Литература

1. Пугачёв В.С. Теория вероятностей и математическая статистика/Учебное 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. Т. 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. Т. 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. Т. 43. № 2. С. 238-244.
5. Лапко А.В., Лапко В.А. Проверка гипотезы о независимости двухмерных случайных величин с использованием непараметрического алгоритма распознавания образов // Автометрия. 2021. Т. 57. № 2. С. 41-48.
6. Parzen E. On estimation of a probability density function and mode // Annals of Mathematical Statistics. 1962. Vol. 33. No. 3. P. 1065-1076.
7. Епанечников В.А. Непараметрическая оценка многомерной плотности вероятности // Теория вероятности и ее применения. 1969. Т. 14. № 1. С. 156-161.
8. Лапко А.В., Медведев А.В., Тишина Е.А. К оптимизации непараметрических оценок // Сб. научных трудов «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 // Scandinavian Journal of Statistics. 1982. No. 9. P. 65-78.
10. Bowman A.W. A comparative study of some kernel-based non-parametric density estimators // Journal of Statistical Computation and Simulation. 1982. Vol. 21. P. 313-327.
11. Hall P. Large-sample optimality of least squares cross-validation in density estimation // Annals of Statistics. 1983. Vol. 11. P. 1156-1174.
12. Jiang M., Provost S.B. A hybrid bandwidth selection methodology for kernel density estimation // Journal of Statistical Computation and Simulation. 2014. Vol. 84, No. 3. P. 614-627.
13. Dutta S. Cross-validation Revisited // Communications in Statistics - Simulation and Computation. 2016. Vol. 45. No. 2. P. 472-490.
14. Heidenreich N.-B., Schindler A., Sperlich S. Bandwidth selection for kernel density estimation: a review of fully automatic selectors // AStA Advances in Statistical Analysis. 2013. Vol. 97. No. 4. P. 403-433.
15. Li Q., Racine J.S. Nonparametric Econometrics: Theory and Practice. Princeton: Princeton University Press. 2007. 768 p.
16. Лапко А.В., Лапко В.А. Методика быстрого выбора коэффициентов размытости в непараметрическом классификаторе, соответствующем критерию максимума апостериорной вероятности // Автометрия. 2019. Т. 55. №6. С. 76-86.
17. Лапко А.В., Лапко В.А. Модифицированный алгоритм быстрого определения коэффициента размытости ядерной оценки плотности вероятности // Автометрия. 2020. Т. 56. № 6. С. 11-18.
18. Scott D.W. Multivariate Density Estimation: Theory, Practice, and Visualization. New Jersey: John Wiley & Sons. 2015. 384 p.
19. Sheather S.J. Density Estimation // Statistical Science. 2004. Vol. 19. № 4. P. 588-597.
20. Silverman B.W. Density Estimation for Statistics and Data Analysis. London: Chapman and Hall. 1986. 175 p.
21. Лапко А.В., Лапко В.А. Оценивание нелинейного функционала от плотности вероятности при оптимизации непараметрических решающих функций // Измерительная техника. 2021. №1. С. 14-20.
22. Лапко А.В., Лапко В.А. Выбор оптимального количества интервалов дискретизации области значений двухмерной случайной величины // Измерительная техника. 2016. № 2. С. 14-17.
23. Лапко А.В., Лапко В. А. Оценивание параметров формулы оптимальной дискретизации области значений двумерной случайной величины // Измерительная техника. 2018. № 5. С. 9-13.

Лапко Александр Васильевич. Доктор технических наук, профессор. Заслуженный деятель науки РФ. Главный научный сотрудник. Институт вычислительного моделирования СО РАН – обособленного подразделения Федерального исследовательского центра «Красноярский научный центр Сибирского отделения Российской академии наук». Профессор кафедры космических средств и технологий. Сибирский государственный университет науки и технологий им. акад. М.Ф. Решетнева. Области исследований: непараметрическая статистика, распознавание образов и анализ изображений, моделирование и оптимизация неопределенных систем, дистанционное зондирование. E-mail: lapko@icm.krasn.ru

Лапко Василий Александрович. Доктор технических наук, профессор. Ведущий научный сотрудник. Институт вычислительного моделирования СО РАН – обособленного подразделения Федерального исследовательского центра «Красноярский научный центр Сибирского отделения Российской академии наук». Заведующий кафедрой космических средств

и технологий. Сибирский государственный университет науки и технологий им. акад. М.Ф. Решетнева. Области исследований: непараметрическая статистика, распознавание образов и анализ изображений, моделирование и оптимизация неопределенных систем, дистанционное зондирование. E-mail: lapko@icm.krasn.ru

Бахтина Анна Владимировна. Заведующий лабораторией дистанционного зондирования. Сибирский государственный университет науки и технологий им. акад. М.Ф. Решетнева. Области исследований: непараметрическая статистика, распознавание образов и анализ изображений, моделирование неопределенных систем, дистанционное зондирование. E-mail: anna-denisyuk@yandex.ru

Comparison of the Methodology for Hypothesis Testing of the Independence of Two-Dimensional Random Variables Based on a Nonparametric Classifier

A. V. Lapko^{1,II}, V. A. Lapko^{1,II}, A. V. Bakhtina^{II}

^IInstitute of Computational Modelling SB RAS, Krasnoyarsk, Russia

^{II}Reshetnev Siberian State University of Science and Technology, Krasnoyarsk, Russia

Annotation. The properties of a new method for hypothesis testing of the independence of random variables based on the use of a nonparametric pattern recognition algorithm corresponding to the maximum likelihood criterion are considered. The estimation of the distribution laws in classes is carried out according to the initial statistical data under the assumption of independence and dependence of the analyzed random variables. Under these conditions, estimates of the probabilities of pattern recognition errors in classes are calculated. According to their minimum value, a decision is made on the independence or dependence of random variables. The results of the proposed method are compared with the Pearson criterion and the Pearson, Spearman and Kendall correlation coefficients. When implementing the Pearson criterion, the formula for optimal discretization of the range of values of a two-dimensional random variable is used. Their effectiveness in complicating the dependence between random variables and changing the volume of initial statistical data is studied by the method of computational experiment.

Keywords: hypothesis testing of the independence of random variables, two-dimensional random variables, nonparametric pattern recognition algorithm, kernel probability density estimation, Pearson criterion, dependent random variables, Pearson, Spearman and Kendall correlation coefficients.

DOI 10.14357/20718594220105

References

1. Pugachev V.S., eds. 2002. Teoriya veroyatnostey i matematicheskaya statistika [Probability theory and mathematical statistics]. Moscow: Fizmatlit Publ. 496 p.
2. Lapko A.V., V.A. Lapko. 2010. Nonparametric algorithms of pattern recognition in the problem of testing a statistical hypothesis on identity of two distribution laws of random variables. Optoelectronics, Instrumentation and Data Processing. 46(6):545–550.
3. Lapko A.V., V.A. Lapko. 2012. Comparison of empirical and theoretical distribution functions of a random variable on the basis of a nonparametric classifier. Optoelectronics, Instrumentation and Data Processing. 48(1):37–41.
4. Lapko A.V., V.A. Lapko. 2019. A technique for testing hypotheses for distributions of multidimensional spectral data using a nonparametric pattern recognition algorithm. Computer Optics. 43(2):238–244.
5. Lapko A.V., V.A. Lapko. 2021. Testing the Hypothesis of the Independence of Two-Dimensional Random Variables Using a Nonparametric Algorithm for Pattern Recognition. Optoelectronics, Instrumentation and Data Processing. 57(2):149–155.
6. Parzen E. 1962. On estimation of a probability density function and mode. Annals of Mathematical Statistics. 33(3):1065–1076.
7. Epanechnikov V.A. 1969. Non-Parametric Estimation of a Multivariate Probability Density. Theory of Probability & Its Applications. 14(1):156–161.
8. Lapko A.V., A.V. Medvedev, E.A. Tishina. 1975. K optimizatsii neparametricheskikh otsenok [To the optimization of nonparametric estimates]. Sbornik nauchnykh trudov "Algoritmy i programmy dlya sistem avtomatizatsii eksperimental'nykh issledovaniy" [Collection of scientific papers "Algorithms and programs for automation systems of experimental research"]. Frunze: Ilim Publ.:105–116.
9. Rudemo M. 1982. Empirical choice of histogram and kernel density estimators. Scandinavian Journal of Statistics. 9(2):65–78.

10. Bowman A.W. 1982. A comparative study of some kernel-based non-parametric density estimators. *Journal of Statistical Computation and Simulation*. 21:313–327.
11. Hall P. 1983. Large-sample optimality of least squares cross-validation in density estimation. *Annals of Statistics*. 11(4):1156–1174.
12. Jiang M., and S.B. Provost. 2014. A hybrid bandwidth selection methodology for kernel density estimation. *Journal of Statistical Computation and Simulation*. 84(3):614–627.
13. Dutta S. 2016. Cross-validation revisited. *Communications in Statistics - Simulation and Computation*. 45(2):472–490.
14. Heidenreich N.B., A. Schindler, S. Sperlich. 2013. Bandwidth selection for kernel density estimation: a review of fully automatic selectors. *AStA Advances in Statistical Analysis*. 97:403–433.
15. Li Q., Racine J.S. 2007. *Nonparametric Econometrics: Theory and Practice*. Princeton: Princeton University Press. 768 p.
16. Lapko A.V., V.A. Lapko. 2019. Method of fast bandwidth selection in a nonparametric classifier corresponding to the a posteriori probability maximum criterion. *Optoelectronics, Instrumentation and Data Processing*. 55(6):597–605.
17. Lapko A.V., V.A. Lapko. 2020. Modified Fast Algorithm for the Bandwidth Selection of the Kernel Density Estimation. *Optoelectronics, Instrumentation and Data Processing*. 56(6):566–572.
18. Scott D.W. 2015. *Multivariate density estimation: Theory, Practice, and Visualization*. New Jersey: John Wiley & Sons. 384 p.
19. Sheather S.J. 2004. Density estimation. *Statistical Science*. 19(4):588–597.
20. Silverman B.W. 1986. *Density estimation for statistics and data analysis*. London: Chapman and Hall. 175 p.
21. Lapko A.V., V.A. Lapko. 2021. Estimation of a nonlinear functional of probability density when optimizing nonparametric decision functions. *Measurement Techniques*. 64(1):13–20.
22. Lapko A.V., and V.A. Lapko. 2016. Selection of the Optimal Number of Intervals Sampling the Region of Values of a Two-Dimensional Random Variable. *Measurement Techniques*. 59(2):122–126.
23. Lapko A.V., V.A. Lapko. 2018. Estimation of parameters of the formula for optimal discretization of the range of values of a two-dimensional random variable. *Measurement Techniques*. 61(5):427–433.

Lapko Alexander V. Doctor of Science in technology, professor. Honored worker of science of the Russian Federation. Chief researcher. Institute of Computational Modelling of the SB RAS – separate subdivision Federal Research Center “Krasnoyarsk Science Center of the Siberian Branch of the Russian Academy of Sciences”. Professor of the Department of space facilities and technologies, Reshetnev Siberian State University of Science and Technology. Research areas: nonparametric statistics; pattern recognition and image analysis; modeling and optimization of uncertain systems; remote sensing. E-mail: lapko@icm.krasn.ru

Lapko Vasily A. Doctor of Science in technology, professor. Leading researcher, Institute of Computational Modelling of the SB RAS – separate subdivision Federal Research Center “Krasnoyarsk Science Center of the Siberian Branch of the Russian Academy of Sciences”. Head of the Department of space facilities and technologies, Reshetnev Siberian State University of Science and Technology. Research areas: nonparametric statistics; pattern recognition and image analysis; modeling of uncertain systems; remote sensing. E-mail: valapko@yandex.ru

Bakhtina Anna V. Head of the Remote Sensing Laboratory, Reshetnev Siberian State University of Science and Technology. Research areas: nonparametric statistics; pattern recognition and image analysis; modeling of uncertain systems; remote sensing. E-mail: anna-denisjuk@yandex.ru