Непараметрический алгоритм распознавания образов в задаче проверки гипотезы о независимости случайных величин

И.В. Зеньков [1,3], А.В. Лапко [2,3], В.А. Лапко [2,3], Е.В. Кирюшина [1], В.Н. Вокин [1]
[1] Сибирский федеральный университет, 660041, г. Красноярск, Россия, пр. Свободный, д. 79, стр. 3;
[2] Институт вычислительного моделирования СО РАН,
660036, Россия, г. Красноярск, Академгородок, д. 50, стр. 44;
[3] Сибирский государственный университет науки и технологий имени академика М.Ф. Решетнева,
660037, г. Красноярск, пр. «Красноярский рабочий», д. 31

Аннотация

Предлагается новая методика проверки гипотезы о независимости многомерных случайных величин. Рассматриваемая методика основывается на использовании непараметрического алгоритма распознавания образов, соответствующего критерию максимального правдоподобия. В отличие от традиционной постановки задачи распознавания образов априори отсутствует обучающая выборка. Исходная информация представляется статистическими данными, которые составляют значения многомерной случайной величины. Законы распределения случайных величин в классах оцениваются по исходным статистическим данным для условий их зависимости и независимости. При выборе оптимальных коэффициентов размытости непараметрических оценок плотностей вероятностей ядерного типа в качестве критерия используется минимум их среднеквадратических отклонений. Вычисляются оценки вероятности ошибки распознавания образов в классах. По минимальному значению оценок вероятностей ошибок распознавания образов принимается решение о независимости либо зависимости случайных величин. Разработанная методика используется при анализе спектральных данных дистанционного зондирования.

Ключевые слова: проверка гипотезы о независимости случайных величин, многомерные случайные величины, распознавание образов, непараметрическая оценка плотности вероятности, коэффициенты размытости ядерных функций, критерий Колмогорова–Смирнова, спектральные данные дистанционного зондирования.

Цитирование: **Зеньков, И.В**. Непараметрический алгоритм распознавания образов в задаче проверки гипотезы о независимости случайных величин / И.В. Зеньков, А.В. Лапко, В.А. Лапко, Е.В. Кирюшина, В.Н. Вокин // Компьютерная оптика. – 2021. – Т. 45, № 5. – С. 767-772. – DOI: 10.18287/2412-6179-CO-871.

Citation: Zenkov IV, Lapko AV, Lapko VA, Kiryushina EV, Vokin VN. Nonparametric pattern recognition algorithm for testing a hypothesis of the independence of random variables. Computer Optics 2021; 45(5): 767-772. DOI: 10.18287/2412-6179-CO-871.

Введение

Сведения о зависимости либо независимости случайных величин являются необходимым условием синтеза эффективных алгоритмов обработки информации и принятия решений. В работе [1] исследованы свойства непараметрической оценки плотности вероятности типа Розенблатта–Парзена независимых случайных величин. Установлено, что наличие априорных сведений о независимости случайных величин позволяет повысить аппроксимационные свойства непараметрической оценки их плотности вероятности по сравнению с ядерной статистикой для зависимых случайных величин. Данное преимущество возрастает с увеличением размерности случайных величин. Полученные результаты подтверждаются при исследовании асимптотических свойств непараметрической оценки уравнения разделяющей поверхности в двухальтернативной задаче распознавания образов [2].

Традиционная методика проверки гипотезы о независимости случайных величин основана на использовании универсального χ – критерия К. Пирсона. Однако его формирование содержит трудно формализуемый этап разбиения области значений случайных величин на многомерные интервалы [3]. Поэтому возникает задача разработки новой методики проверки рассматриваемой гипотезы, обеспечивающей обход проблемы декомпозиции области значений случайных величин. Подобная задача решается при проверке гипотезы о тождественности законов распределения случайных величин на основе использования непараметрического алгоритма распознавания образов [4]. Показана возможность её замены на задачу проверки гипотезы о равенстве ошибки распознавания образов определённому пороговому значению. Обучающая выборка при синтезе непараметрического алгоритма распознавания образов формируется по статистическим данным, характеризующих законы распределения сравниваемых случайных величин.

Цель рассматриваемого исследования состоит в развитии предложенного подхода на задачу проверки гипотезы о независимости многомерных случайных величин с использованием непараметрического алгоритма распознавания образов, соответствующего критерию максимального правдоподобия.

Постановка задачи

Пусть имеется выборка $V=(x_v^i, v=\overline{1,k}, i=\overline{1,n})$ объёма n, сформированная из независимых наблюдений многомерной случайной величины $x=(x_v, v=\overline{1,k})$. Наблюдения x извлекаются из генеральных совокупностей, характеризуемых неизвестными плотностями вероятности

$$\prod_{v=1}^{k} p(x_v) \text{ либо } p\left(x_v, v=\overline{1,k}\right).$$

Необходимо по статистическим данным V проверить гипотезу

$$H_0: \; p\left(x_v, v=\overline{1,k}\right) \equiv \prod_{v=1}^{k} p(x_v) \tag{1}$$

о независимости случайных величин $x=(x_v, v=\overline{1,k})$.

Модификация непараметрического алгоритма распознавания образов

Для проверки гипотезы H_0 (1) будем решать двухальтернативную задачу распознавания образов. Родственность задач распознавания образов и проверки гипотез отмечалась в работах Л.Л. Леймана (1959), В.С. Пугачёва (1979). Под классами Ω_1, Ω_2 понимаются области определения плотностей вероятностей

$$\prod_{v=1}^{k} p(x_v), \; p\left(x_v, v=\overline{1,k}\right).$$

В этих условиях байесовское решающее правило, соответствующее критерию максимального правдоподобия, имеет вид

$$m(x): \begin{cases} x\in\Omega_1, \text{ если } p\left(x_v, v=\overline{1,k}\right) < \prod_{v=1}^{k} p(x_v) \\ x\in\Omega_2, \text{ если } p\left(x_v, v=\overline{1,k}\right) > \prod_{v=1}^{k} p(x_v). \end{cases} \tag{2}$$

В отличие от традиционной постановки задачи распознавания образов при синтезе решающего правила (2) в условиях исходной неопределённости отсутствует обучающая выборка. Оценивание плотностей вероятностей

$$\prod_{v=1}^{k} p(x_v), \; p\left(x_v, v=\overline{1,k}\right)$$

осуществляется по выборке V. Для этого используются непараметрические статистики типа Розенблатта–Парзена [5, 6]

$$\overline{p}\left(x_v, v=\overline{1,k}\right) = \frac{1}{n\prod_{v=1}^{k} c_v} \sum_{i=1}^{n} \prod_{v=1}^{k} \Phi\left(\frac{x_v - x_v^i}{c_v}\right), \tag{3}$$

$$\prod_{v=1}^{k} \overline{p}(x_v) = \frac{1}{n^k \prod_{v=1}^{k} c_v} \prod_{v=1}^{k} \sum_{i=1}^{n} \Phi\left(\frac{x_v - x_v^i}{c_v}\right). \tag{4}$$

В статистиках (3), (4) ядерные функции $\Phi(u_v)$ удовлетворяют условиям:

$$\Phi(u_v) = \Phi(-u_v), 0 \le \Phi(u_v) < \infty, \int_{-\infty}^{+\infty} \Phi(u_v) du_v = 1,$$

$$\int_{-\infty}^{+\infty} u^m \Phi(u_v) du_v < \infty, 0 \le m < \infty, v = \overline{1,k}.$$

Значения коэффициентов размытости c_v ядерных функций убывают с ростом объёма n выборки статистических данных V. Тогда с учётом выражений (2)–(4) непараметрическое решающее правило классификации случайных величин $x=(x_v, v=\overline{1,k})$ запишется как

$$\overline{m}(x): \begin{cases} x\in\Omega_1, \text{ если } \overline{p}\left(x_v, v=\overline{1,k}\right) < \prod_{v=1}^{k} \overline{p}(x_v) \\ x\in\Omega_2, \text{ если } \overline{p}\left(x_v, v=\overline{1,k}\right) > \prod_{v=1}^{k} \overline{p}(x_v). \end{cases} \tag{5}$$

В приведённой модификации непараметрического алгоритма распознавания образов (5) оптимальные коэффициенты размытости c_v, $v=\overline{1,k}$ ядерных функций оценок плотностей вероятностей

$$\overline{p}\left(x_v, v=\overline{1,k}\right), \; \prod_{v=1}^{k} \overline{p}(x_v)$$

будем выбирать на основе анализа их аппроксимационных свойств. Например, для определения оптимального коэффициента размытости c_v ядерных функций непараметрической оценки плотности вероятности $\overline{p}(x_v)$ рассмотрим критерий

$$W(c_v) = \int \left(\overline{p}(x_v) - p(x_v)\right)^2 dx_v, \tag{6}$$

который характеризует меру близости между $\overline{p}(x_v)$ и $p(x_v)$.

Преобразуем с учётом непараметрической оценки плотности вероятности $\overline{p}(x_v)$ выражение (6)

$$W(c_v) = \frac{1}{n^2 c_v^2} \sum_{j=1}^{n} \sum_{i=1}^{n} \int \Phi\left(\frac{x_v - x_v^j}{c_v}\right) \Phi\left(\frac{x_v - x_v^i}{c_v}\right) dx_v - $$

$$- \frac{2}{n c_v} \sum_{i=1}^{n} \int \Phi\left(\frac{x_v - x_v^i}{c_v}\right) p(x_v)\, dx_v + \int p^2(x_v)\, dx_v.$$

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

$$\overline{b}(c_v) = -\frac{2}{n^2 c_v} \sum_{j=1}^{n} \sum_{\substack{i=1 \\ i\ne j}}^{n} \Phi\left(\frac{x_v^j - x_v^i}{c_v}\right). \tag{7}$$

При выполнении условия $i\ne j$ статистика (7) является несмещённой оценкой $b(c_v)$.

Тогда оптимальные значения $\overline{c}_v$ будем находить путём минимизации критерия

$$\overline{W}(c_v)=\frac{1}{n^2c_v^2}\sum_{j=1}^{n}\sum_{i=1}^{n}\int\Phi\left(\frac{x_v-x_v^j}{c_v}\right)\Phi\left(\frac{x_v-x_v^i}{c_v}\right)dx_v-$$
$$-\frac{2}{n^2c_v}\sum_{j=1}^{n}\sum_{\substack{i=1\\ i\neq j}}^{n}\Phi\left(\frac{x_v^j-x_v^i}{c_v}\right). \quad (8)$$

Возможность использования критерия (8) для выбора оптимальных коэффициентов размытости в $\overline{p}(x_v)$ заключается в том, что статистическая оценка $\overline{b}(c_v)$ имеет значительно большую скорость сходимости к $b(c_v)$ с ростом n, чем $\overline{p}(x_v)$ к $p(x_v)$.

По аналогии с выражением (8) нетрудно определить критерий выбора оптимальных коэффициентов размытости статистики $\overline{p}\left(x_v, v=\overline{1,k}\right)$ (3).

Впервые подход к оптимизации непараметрической оценки плотности вероятности типа Розенблатта–Парзена по коэффициенту размытости ядерных функций из условия минимума статистической оценки среднеквадратического отклонения $\overline{p}(x_v)$ от $p(x_v)$ был предложен в работе [7]. Эта методика позднее была повторена в статьях [8, 9] и является актуальной до настоящего времени [10 – 13]. Исследованы её свойства при использовании ядерных функций, соответствующих нормальному закону [13]. В этих условиях значительно упрощаются вычисления критерия оптимизации $\overline{p}(x_v)$ по значениям $c_v(n)$. Выбор оптимальных значений коэффициентов размытости ядерных функций, соответствующих максимуму функции правдоподобия, рассмотрен в работах [14, 15].

В отличие от приведённых выше методов оптимизации непараметрического решающего правила (5) по коэффициентам размытости ядерных функций c_v, $v=\overline{1,k}$, будем полагать, что в статистиках (3), (4) значение $c_v=c\,\overline{\sigma}_v$. Здесь $\overline{\sigma}_v$ – оценка средних квадратических отклонений случайной величины x_v в выборке V. Данное утверждение является очевидным, так как большей длине интервала значений соответствует больший коэффициент размытости c_v ядерных функций $\Phi(u_v)$, $v=\overline{1,k}$. Предложенная методика выбора коэффициентов размытости ядерных функций использовалась при формировании быстрых процедур оптимизации непараметрических оценок плотности вероятности [16 – 19].

Значения оценок средних квадратических отклонений $\overline{\sigma}_v$ определяются по статистическим данным выборки V

$$\overline{\sigma}_v=\left(\frac{1}{n-1}\sum_{i=1}^{n}\left(x_v^i-\overline{x}_v\right)^2\right)^{1/2},\ v=\overline{1,k}.$$

Здесь $\overline{x}_v$ – среднее значение случайной величины x_v, которое вычисляется по выборке V.

Поэтому появляется возможность оптимизацию непараметрического алгоритма распознавания образов (5) проводить лишь по одному параметру c коэффициентов размытости ядерных функций.

Методика проверки гипотезы о независимости случайных величин

Предлагаемая методика основана на выполнении следующих действий.

1. На основе решающего правила (5) и ядерных оценок плотностей вероятностей осуществить синтез модификации непараметрического алгоритма распознавания образов. В качестве информации используется выборка V значений многомерной случайной величины $x=(x_v, v=\overline{1,k})$ объёма n. Оптимальные коэффициенты размытости ядерных функций непараметрических оценок плотностей вероятностей (3), (4) определяются из условия минимума критерия типа (8).

2. Оценить значения вероятностей ошибок распознавания образов. Определим оценки вероятностей ошибок распознавания образов $\overline{\rho}_1$, $\overline{\rho}_2$ решающим правилом (5) по исходным статистическим данным V при оптимальных коэффициентах размытости ядерных функций статистик

$$\prod_{v=1}^{k}\overline{p}(x_v),\ \overline{p}\left(x_v, v=\overline{1,k}\right).$$

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

$$\overline{\rho}_t=\frac{1}{n}\sum_{j=1}^{n}1\left(\delta(j),\overline{\delta}(j)\right),\ t=1,2,$$

где $\delta(j)=t$ – указания типа $x_t\in\Omega_t$;

$$\overline{\delta}(j)=\begin{cases}t, \text{ если } x^j\in\Omega_t\\ 0, \text{ если } x^j\notin\Omega_t\end{cases} -$$

«решение» алгоритма (5) о принадлежности ситуации x^j к одному из классов Ω_t, $t=1,2$.

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

Индикаторная функция определяется выражением

$$1\left(\delta(j),\overline{\delta}(j)\right)=\begin{cases}0, \text{если } \delta(j)=\overline{\delta}(j)\\ 1, \text{если } \delta(j)\neq\overline{\delta}(j).\end{cases}$$

Обозначим через $\overline{\overline{\rho}}_t$ минимальное значение оценки вероятности ошибки распознавания образов в предположении, что элементы выборки V принадлежат классу Ω_t, $t=1,2$.

3. Сравнить значения $\overline{\overline{\rho}}_1$, $\overline{\overline{\rho}}_2$ в предположении, что элементы выборки V принадлежат классам Ω_1, Ω_2 соответственно.

Тогда гипотеза H_0 справедлива, если $\overline{\overline{\rho}}_1$ меньше $\overline{\overline{\rho}}_2$. В противном случае при $\overline{\overline{\rho}}_2$ меньше $\overline{\overline{\rho}}_1$ случайные величины x_1 и x_2 являются зависимыми.

Естественно, что при ограниченных объёмах n выборки V возникает задача доверительного оцени-

вания вероятностей ошибок распознавания образов. Для её решения может использоваться традиционная методика доверительного оценивания вероятностей [3] либо критерий Колмогорова–Смирнова [20].

Например, при использовании критерия Колмогорова–Смирнова отклонение $\overline{D}_{12} = \left|\overline{\rho}_1 - \overline{\rho}_2\right|$ сравнивается с пороговым значением [20]

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

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

Анализ результатов вычислительных экспериментов

Рассмотрим применение разработанного метода проверки гипотез о независимости случайных величин при анализе данных дистанционного зондирования.

Исследуемая территория соответствует горной лесотундре, расположенной в западной части Алтае-Саянского региона (50°03’ северной широты, 85°15’ восточной долготы) на высоте 2273 метра над уровнем моря. Исходная информация формировалась по фрагменту спутниковой съемки Worldview-2 (https://www.satimagingcorp.com/satellite-sensors/worldview-2) с пространственным разрешением 0,6 метра. Размер фрагмента составляет 162×192 пикселя, а его площадь равна 1,1 га. Каждый пиксель характеризовался четырьмя спектральными каналами: синий (x_1), зеленый (x_2), красный (x_3), ближний инфракрасный (x_4). Объектом исследования являются элементы земной поверхности, включающие кедровый стланик (Pinus sibirica) в виде полос с тенями. Просматриваются участки с травяно-кустарниковым покровом и выходами горных пород. На анализируемом участке ранее проводились исследования пространственно-временной динамики роста кедра и кедрового стланика в изменяющихся климатических условиях [21].

Количественные характеристики законов распределения спектральных признаков x_v, $v = \overline{1, 4}$ при объёме статистических данных $n = 31104$ приведены в табл. 1.

Табл. 1. Количественные характеристики законов распределения спектральных признаков, характеризующих элементы территории горной лесотундры западной части Алтае-Саянского региона

Спектральные признаки	$\overline{x}_v$	$\overline{\sigma}_v$	$\overline{\delta}_v$	$\overline{\gamma}_v$
x_1	187,27	27,396	0,508	–0,993
x_2	240,37	53,178	0,518	–0,844
x_3	140,98	45,363	0,537	–0,857
x_4	412,81	95,992	0,510	–0,520

Обозначения количественных характеристик законов распределения случайных величин x_v, $v = \overline{1, 4}$: $\overline{x}_v$ – среднее значение; $\overline{\sigma}_v$ – среднее квадратическое отклонение; $\overline{\delta}_v$ – коэффициент контрэксцесса; $\overline{\gamma}_v$ – коэффициент асимметрии.

Табл. 2. Результаты проверки гипотез о независимости спектральных признаков по предложенной методике

Значения	Спектральные признаки					
	x_1, x_2	x_1, x_3	x_1, x_4	x_2, x_3	x_2, x_4	x_3, x_4
$\overline{\rho}_1$	0,99	0,959	0,844	0,971	0,856	0,826
$\overline{\rho}_2$	0,01	0,041	0,156	0,029	0,144	0,174
$\overline{r}$	0,983	0,964	0,808	0,977	0,877	0,806

Здесь значение $\overline{\rho}_1$ определяет оценку вероятности ошибки отнесения ситуаций из исходной выборки V к классу независимых случайных величин в соответствии с алгоритмом (5), а $\overline{\rho}_2$ – к классу их зависимых значений. Символом $\overline{r}$ обозначена оценка коэффициента корреляции между анализируемыми спектральными признаками.

Полученные результаты проверки гипотез о независимости спектральных признаков, представленные в табл. 2, подтверждают эффективность предложенной методики. Установлено, что исследуемые спектральные признаки являются зависимыми, так как для всех их парных сочетаний значения $\overline{\rho}_2$ значительно меньше $\overline{\rho}_1$. Например, для спектральных признаков x_1, x_2 значения $\overline{\rho}_1 = 0,99$, а $\overline{\rho}_2 = 0,01$. В этих условиях соотношение $\overline{\rho}_2 < \overline{\rho}_1$ выполняется для большинства элементов выборки $(x_1^i, x_2^i, i = \overline{1, n})$ и справедливо неравенство $\overline{p}(x_1, x_2) > \overline{p}(x_1)\,\overline{p}(x_2)$. Следовательно, предположение о зависимости спектральных признаков x_1, x_2 выполняется. Причём обнаруженная зависимость близка к линейной, что подтверждается достаточно большим значением оценки коэффициента корреляции $\overline{r} = 0,983$. Полученный вывод свойственен также другим сочетаниям спектральных признаков и особо характерен для пар: (x_1, x_2), (x_1, x_3), (x_2, x_3) (табл. 2). Можно определить группу спектральных признаков (x_1, x_2, x_3) с высоким уровнем линейных зависимостей, для которых значения $\overline{\rho}_1 \in [0,959;\ 0,99]$, $\overline{\rho}_2 \in [0,01;\ 0,041]$, а оценка коэффициента корреляции $\overline{r} \in [0,964;\ 0,977]$.

При объёме исходных статистических данных $n = 31104$ и заданном риске $\beta = 0,05$ отвергнуть гипотезу $\overline{H}_0 : \rho_1 = \rho_2$ значения вероятностей ошибок распознавания образов в предположении зависимости и независимости спектральных признаков отличаются достоверно. В этих условиях при использовании критерия Колмогорова–Смирнова для всех анализируемых ситуаций $\overline{D}_{12} = \left|\overline{\rho}_1 - \overline{\rho}_2\right|$ значительно больше порогового значения $D_\beta = 0,01$. В соответствии с предлагаемой методикой гипотеза о независимости случайных величин H_0 отвергается, а при выполнении соотношений $\overline{D}_{12} > D_\beta = 0,01$ и $\overline{\rho}_2 < \overline{\rho}_1$ анализируемые спектральные признаки являются зависимыми.

Заключение

Предложенная методика проверки гипотезы о независимости случайных величин обеспечивает обход проблемы декомпозиции области значений случайных величин на многомерные интервалы, которая свойственна критерию Пирсона. Для проверки гипотезы используется непараметрический алгоритм распознавания образов, соответствующий критерию максимального правдоподобия. Каждый класс определяется многомерной плотностью вероятности в предположении независимости либо зависимости случайных величин. Выбор коэффициентов размытости ядерных оценок плотностей вероятностей случайных величин в классах осуществляется из условия минимума их среднеквадратических отклонений. Используя исходные статистические данные, вычисляются оценки вероятностей ошибок распознавания ситуаций, принадлежащих введённым классам. По их минимальному значению принимается решение о независимости либо зависимости случайных величин.

Существует линейная зависимость между парными сочетаниями спектральных признаков x_1, x_2, x_3, x_4, что подтверждается большими значениями оценок коэффициентов корреляции и результатами проверки достоверности полученных выводов. Группа признаков (x_1, x_2, x_3), формируемая спектральными каналами (синий (x_1), зелёный (x_2), красный (x_3)), характеризуются наиболее высокими показателями зависимости. Подобные сведения являются необходимыми при выборе признаков в задаче синтеза эффективных алгоритмов принятия решений.

Перспективным исследованием в данном направлении является применение предлагаемой методики при формировании наборов независимых случайных величин, что позволит упростить задачу синтеза эффективных алгоритмов обработки информации.

Благодарности

Исследование выполнено при финансовой поддержке РФФИ, Правительства Красноярского края и Красноярского краевого фонда науки в рамках научного проекта № 20-41-240001.

Литература

1. **Лапко, А.В.** Свойства непараметрической оценки многомерной плотности вероятности независимых случайных величин / А.В. Лапко, В.А. Лапко // Информатика и системы управления. – 2012. – Т. 31, № 1. – С. 166-174.
2. **Лапко, А.В.** Свойства непараметрической решающей функции при наличии априорных сведений о независимости признаков классифицируемых объектов / А.В. Лапко, В.А. Лапко // Автометрия. – 2012. – Т. 48, № 4. – С. 112-119.
3. **Пугачёв, В.С.** Теория вероятностей и математическая статистика: учебное пособие / В.С. Пугачёв. – М: Физматлит, 2002. – 496 с.
4. **Лапко, А.В.** Методика проверки гипотез о распределениях многомерных спектральных данных с использованием непараметрического алгоритма распознавания образов / А.В. Лапко, В.А. Лапко // Компьютерная оптика. – Т. 2019. – Т. 43, № 2. – С. 238-244. – DOI: 10.18287/2412-6179-2019-43-2-238-244.
5. **Parzen, E.** On estimation of a probability density function and mode / E. Parzen // Annals of Mathematical Statistics. – 1962. – Vol. 33, Issue 3. – P. 1065-1076. – DOI: 10.1214/aoms/1177704472.
6. **Епанечников, В.А.** Непараметрическая оценка многомерной плотности вероятности / В.А. Епанечников // Теория вероятности и ее применения. – 1969. – Т. 14, № 1. – С. 156-161.
7. **Лапко, А.В.** К оптимизации непараметрических оценок / А.В. Лапко, А.В. Медведев, Е.А. Тишина // Сборник научных трудов «Алгоритмы и программы для систем автоматизации экспериментальных исследований». – Фрунзе: Илим, 1975. – С. 105-116.
8. **Rudemo, M.** Empirical choice of histogram and kernel density estimators / M. Rudemo // Scandinavian Journal of Statistics. – 1982. – Vol. 9, No. 2 – P. 65-78.
9. **Hall, P.** Large sample optimality of least squares cross-validation in density estimation / P. Hall // Annals of Statistics. – 1983. – Vol. 11, No. 4. – P. 1156-1174.
10. **Jiang, M.** A hybrid bandwidth selection methodology for kernel density estimation / M. Jiang, S.B. Provost // Journal of Statistical Computation and Simulation. – 2014. – Vol. 84, Issue 3. – P. 614-627. – DOI: 10.1080/00949655.2012.721366.
11. **Dutta, S.** Cross-validation revisited / S. Dutta // Communications in Statistics – Simulation and Computation. – 2016. – Vol. 45, Issue 2. – P. 472-490. – DOI: 10.1080/03610918.2013.862275.
12. **Heidenreich, N.B.** Bandwidth selection for kernel density estimation: a review of fully automatic selectors / N.B. Heidenreich, A. Schindler, S. Sperlich // AStA Advances in Statistical Analysis. – 2013. – Vol. 97. – P. 403-433. – DOI: 10.1007/s10182-013-0216-y.
13. **Li, Q.** Nonparametric econometrics: Theory and practice / Q. Li, J.S. Racine. – Princeton: Princeton University Press, 2007. – 768 p.
14. **Duin, R.** On the choice of smoothing parameters for parzen estimators of probability density functions / R. Duin // IEEE Transactions on Computers. – 1976. – Vol. C-25, Issue 11. – P. 1175-1179. – DOI: 10.1109/TC.1976.1674577.
15. **Botev, Z.I.** Non-asymptotic bandwidth selection for density estimation of discrete data / Z.I. Botev, D.P. Kroese // Methodology and Computing in Applied Probability. – 2008. – Vol. 10, Issue 3. – P. 435-451. – DOI: 10.1007/s11009-007-9057-z.
16. **Лапко, А.В.** Методика быстрого выбора коэффициентов размытости в непараметрическом классификаторе, соответствующем критерию максимума апостериорной вероятности / А.В. Лапко, В.А. Лапко // Автометрия. – 2019. – Т. 55, № 6. – С. 76-86. – DOI: 10.15372/AUT20190610.
17. **Scott, D.W.** Multivariate density estimation: Theory, practice, and visualization / D.W. Scott. – New Jersey: John Wiley & Sons, 2015. – 384 p.
18. **Sheather, S.J.** Density estimation / S.J. Sheather // Statistical Science. – 2004. – Vol. 19, Issue 4. – P. 588-597. – DOI: 10.1214/088342304000000297.
19. **Silverman, B.W.** Density estimation for statistics and data analysis / B.W. Silverman. – London: Chapman and Hall, 1986. – 175 p.
20. **Шаракшанэ, А.С.** Сложные системы / А.С. Шаракшанэ, И.Г. Железнов, В.А. Ивницкий. – М.: Высшая школа, 1977. – 248 с.
21. **Kharuk, V.I**. Tree wave migration across an elevation gradient in the Altai Mountains, Siberia / V.I. Kharuk, S.T. Im, M.L. Dvinskaya, K.J. Ranson, I.A. Petrov // Journal of Mountain Science. – 2017. – Vol. 14, No. 3. – P. 442-452. – DOI: 10.1007/s11629-016-4286-7.

Сведения об авторах

Зеньков Игорь Владимирович, 1963 года рождения, в 1985 г. окончил Красноярский институт цветных металлов по специальности «Технология и комплексная механизация открытой разработки месторождений полезных ископаемых», доктор технических наук, профессор, профессор кафедры систем автоматики, автоматизированного управления и проектирования Сибирского федерального университета, ведущий научный сотрудник Красноярского филиала Федерального исследовательского центра информационных и вычислительных технологий. Область научных интересов: решение задач горнодобывающей промышленности с использованием ресурсов дистанционного зондирования; информационное обеспечение мониторинга технологических, логистических параметров предприятий горной промышленности; дистанционное зондирование.
E-mail: *zenkoviv@mail.ru* .

Лапко Александр Васильевич, 1949 года рождения, в 1971 году окончил Фрунзенский политехнический институт по специальности «Автоматика и телемеханика», доктор технических наук, профессор, заслуженный деятель науки РФ, главный научный сотрудник Института вычислительного моделирования Сибирского отделения Российской академии наук, профессор кафедры космических средств и технологий Сибирского государственного университета науки и технологий имени академика М.Ф. Решетнева. Область научных интересов: непараметрическая статистика; распознавание образов и анализ изображений; моделирование и оптимизация неопределённых систем, дистанционное зондирование. E-mail: *lapko@icm.krasn.ru* .

Лапко Василий Александрович, 1974 года рождения, в 1996 году окончил Красноярский государственный технический университет по специальности «Управление и информатика в технических системах», доктор технических наук, профессор, ведущий научный сотрудник Института вычислительного моделирования Сибирского отделения Российской академии наук, заведующий кафедрой космических средств и технологий Сибирского государственного университета науки и технологий имени академика М.Ф. Решетнева. Область научных интересов: непараметрическая статистика; распознавание образов и анализ изображений; моделирование неопределённых систем, дистанционное зондирование. E-mail: *valapko@yandex.ru* .

Кирюшина Елена Васильевна, 1963 года рождения, в 1985 г. окончила Красноярский институт цветных металлов по специальности «Технология и комплексная механизация открытой разработки месторождений полезных ископаемых», кандидат технических наук, доцент, доцент кафедры открытых горных работ Сибирского федерального университета. Область научных интересов: решение задач горнодобывающей промышленности с использованием ресурсов дистанционного зондирования; информационное обеспечение мониторинга технологических, логистических параметров предприятий горной промышленности; дистанционное зондирование. E-mail: *kiryushinaev@mail.ru* .

Вокин Владимир Николаевич, 1954 года рождения, в 1976 г. окончил Красноярский институт цветных металлов по специальности «Технология и комплексная механизация открытой разработки месторождений полезных ископаемых», кандидат технических наук, доцент, профессор кафедры открытых горных работ Сибирского федерального университета. Область научных интересов: решение задач горнодобывающей промышленности с использованием ресурсов дистанционного зондирования; информационное обеспечение мониторинга технологических, логистических параметров предприятий горной промышленности; дистанционное зондирование. E-mail: *vokin@krasmail.ru* .

ГРНТИ: 28.23.15
Поступила в редакцию 29 января 2021 г. Окончательный вариант – 26 мая 2021 г.

Nonparametric pattern recognition algorithm for testing a hypothesis of the independence of random variables

I.V. Zenkov [1,3], A.V. Lapko [2,3], V.A. Lapko [2,3], E.V. Kiryushina [1], V.N. Vokin [1]
[1] Siberian Federal University,
660041, Krasnoyarsk, Russia, Svobodny Av. 79,
[2] Institute of Computational Modelling SB RAS,
660036, Krasnoyarsk, Russia, Akademgorodok 50,
[3] Reshetnev Siberian State University of Science and Technology,
660037, Krasnoyarsk, Russia, Krasnoyarsky Rabochy Av. 31

Abstract

A new method for testing a hypothesis of the independence of multidimensional random variables is proposed. The technique under consideration is based on the use of a nonparametric pattern recognition algorithm that meets a maximum likelihood criterion. In contrast to the traditional formulation of the pattern recognition problem, there is no a priori training sample. The initial information is represented by statistical data, which are made up of the values of a multivariate random variable. The distribution laws of random variables in the classes are estimated according to the initial statistical data for the conditions of their dependence and independence. When selecting optimal bandwidths for nonparametric kernel-type probability density estimates, the minimum standard deviation is used as a criterion. Estimates of the probability of pattern recognition error in the classes are calculated. Based on the minimum value of the estimates of the probabilities of pattern recognition errors, a decision is made on the independence or dependence of the random variables. The technique developed is used in the spectral analysis of remote sensing data.

Keywords: testing a hypothesis of the independence of random variables, multidimensional random variables, pattern recognition, nonparametric probability density estimation, bandwidths of kernel functions, Kolmogorov–Smirnov criterion, spectral analysis of remote sensing data.

Citation: Zenkov IV, Lapko AV, Lapko VA, Kiryushina EV, Vokin VN. Nonparametric pattern recognition algorithm for testing a hypothesis of the independence of random variables. Computer Optics 2021; 45(5): 767-772. DOI: 10.18287/2412-6179-CO-871.

Acknowledgements: The research was funded by the Russian Foundation for Basic Research, government of Krasnoyarsk Territory, and Krasnoyarsk Regional Science Foundation under project No. 20-41-240001.

References

[1] Lapko AV, Lapko VA. Properties of nonparametric estimates of multidimensional probability density of independent random variables [In Russian]. Informatika i Sistemy Upravleniya 2012; 31(1): 166-174.

[2] Lapko AV, Lapko VA. Properties of the nonparametric decision function with a priori information on independence of attributes of classified objects. Optoelectronics, Instrumentation and Data Processing 2012; 48(4): 416-422. DOI: 10.3103/S8756699012040139.

[3] Pugachev VS. Probability theory and mathematical statistics: textbook [In Russian]. Moscow: "Fizmatlit" Publisher; 2002.

[4] Lapko AV, Lapko VA. A technique for testing hypotheses for distributions of multidimensional spectral data using a nonparametric pattern recognition algorithm. Computer Optics 2019; 43(2): 238-244. DOI: 10.18287/2412-6179-2019-43-2-238-244

[5] Parzen E. On estimation of a probability density function and mode. Ann Math Statistic 1962; 33(3): 1065-1076. DOI: 10.1214/aoms/1177704472.

[6] Epanechnikov VA. Non-parametric estimation of a multivariate probability density. Theory Probab its Appl 1969; 14(1): 153-158. DOI: 10.1137/1114019.

[7] Lapko AV, Medvedev AV, Tishina EA. To the optimization of nonparametric estimates [In Russian]. Collection of scientific papers "Algorithms and programs for automation systems of experimental research" (Frunze: Ilim) 1975: 105-116.

[8] Rudemo M. Empirical choice of histogram and kernel density estimators. Scand Stat Theory Appl 1982; 9(2): 65-78.

[9] Hall P. Large sample optimality of least squares cross-validation in density estimation. Ann Stat 1983; 11(4): 1156-1174.

[10] Jiang M, Provost SB. A hybrid bandwidth selection methodology for kernel density estimation. J Stat Comput Simul 2014; 84(3): 614-627. DOI: 10.1080/00949655.2012.721366.

[11] Dutta S. Cross-validation revisited. Commun Stat Simul Comput 2016; 45(2): 472-490. DOI: 10.1080/03610918.2013.862275.

[12] Heidenreich NB, Schindler A, Sperlich S. Bandwidth selection for kernel density estimation: a review of fully automatic selectors. Adv Stat Anal 2013; 97: 403-433. DOI: 10.1007/s10182-013-0216-y.

[13] Li Q, Racine JS. Nonparametric econometrics: Theory and practice. Princeton: Princeton University Press; 2007.

[14] Duin R. On the choice of smoothing parameters for Parzen estimators of probability density functions. IEEE Trans Comput 1976; C-25(11): 1175-1179. DOI: 10.1109/TC.1976.1674577.

[15] Botev ZI, Kroese DP. Non-asymptotic bandwidth selection for density estimation of discrete data. Methodol Comput Appl Probab 2008; 10(3): 435-451. DOI: 10.1007/s11009-007-9057-z.

[16] Lapko AV, Lapko VA. Method of fast bandwidth selection in a nonparametric classifier corresponding to the a posteriori probability maximum criterion. Optoelectronics, Instrumentation and Data Processing 2019; 55(6): 597-605. DOI: 10.3103/S8756699019060104.
[17] Scott DW. Multivariate density estimation: Theory, practice, and visualization. New Jersey: John Wiley and Sons; 2015.
[18] Sheather SJ. Density estimation. Stat Sci 2004; 19(4): 588-597. DOI: 10.1214/088342304000000297.
[19] Silverman BW. Density estimation for statistics and data analysis. London: Chapman and Hall; 1986.
[20] Sharakshaneh AS, Zheleznov IG, Ivnitskij VA. Complex system [In Russian]. Moscow: "Vysshaya shkola" Publisher; 1977.
[21] Kharuk VI, Im ST, Dvinskaya ML, Ranson KJ, Petrov IA. Tree wave migration across an elevation gradient in the Altai Mountains, Siberia. J Mt Sci 2017; 14(3): 442-452. DOI: 10.1007/s11629-016-4286-7.

Authors' information

Igor Vladimirovich Zenkov (b. 1963), graduated from Krasnoyarsk Institute of Non-ferrous Metals on speciality "Technology and Complex Mechanization of Opencast Mining of Mineral Deposits" in 1985. Doctor of Science in Technology, professor, professor of Automation Systems, Automated Control and Design department at the Siberian Federal University; leading researcher at the Krasnoyarsk branch of the Federal Research Center for Information and Computational Technologies. Research interests: solving problems in the mining industry using remote sensing resources; information support for monitoring technological, logistic parameters of mining enterprises; remote sensing.
E-mail: *zenkoviv@mail.ru* .

Alexander Vasilievich Lapko (b. 1949), graduated from Frunze Polytechnic Institute on speciality "Automation and Telemechanics" in 1971. Doctor of Science in Technology, professor, honored worker of science of the Russian Federation, chief researcher of the Institute of Computational Modeling of the Siberian Branch of the Russian Academy of Sciences; Professor of Space Facilities and Technologies department of the Reshetnev Siberian State University of Science and Technology. Research interests: nonparametric statistics; pattern recognition and image analysis; modeling and optimization of uncertain systems; remote sensing. E-mail: *lapko@icm.krasn.ru* .

Vasiliy Aleksandrovich Lapko (b. 1974), graduated from Krasnoyarsk State Technical University on speciality "Management and Informatics in Technical Systems" in 1996. Doctor of Science in Technology, professor, leading researcher at the Institute of Computational Modeling of the Siberian Branch of the Russian Academy of Sciences; Head of Space Facilities and Technologies department of the Reshetnev Siberian State University of Science and Technology. Research interests: nonparametric statistics; pattern recognition and image analysis; modeling of uncertain systems; remote sensing. E-mail: *valapko@yandex.ru* .

Elena Vasilievna Kiryushina (b. 1963), graduated from Krasnoyarsk Institute of Non-ferrous Metals on speciality "Technology and Complex Mechanization of Opencast Mining of Mineral Deposits" in 1985. Candidate of Science in Technology, associate professor, associate professor of the Open Mining department at the Siberian Federal University. Research interests: solving problems in the mining industry using remote sensing resources; information support for monitoring technological, logistic parameters of mining enterprises; remote sensing. E-mail: *kiryushinaev@mail.ru* .

Vladimir Nikolaevich Vokin (b. 1954), graduated from Krasnoyarsk Institute of Non-ferrous Metals on speciality "Technology and Complex Mechanization of Opencast Mining of Mineral Deposits" in 1976. Candidate of Science in Technology, associate professor, associate professor of Open Mining department at the Siberian Federal University. Research interests: solving problems in the mining industry using remote sensing resources; information support for monitoring technological, logistic parameters of mining enterprises; remote sensing. E-mail: *vokin@krasmail.ru* .

Received January 29, 2021. The final version – May 26, 2021.