УДК 681.513

© 2014 г. **А.В. Лапко**, д-р техн. наук,
В.А. Лапко, д-р техн. наук

(Институт вычислительного моделирования СО РАН,
Сибирский государственный аэрокосмический университет имени
академика М.Ф. Решетнева, Красноярск)

РЕГРЕССИОННАЯ ОЦЕНКА УРАВНЕНИЯ РАЗДЕЛЯЮЩЕЙ ПОВЕРХНОСТИ В ДВУАЛЬТЕРНАТИВНОЙ ЗАДАЧЕ РАСПОЗНАВАНИЯ ОБРАЗОВ*

Исследуется количественная зависимость аппроксимационных свойств регрессионной оценки уравнения разделяющей поверхности в двуальтернативной задаче распознавания образов от методов дискретизации значений случайной величины в классах.

Ключевые слова: уравнение разделяющей поверхности, распознавание образов, регрессионная оценка, асимптотические свойства, методы дискретизации.

Введение

Применение теории классификации и методов непараметрической статистики – одно из перспективных направлений исследования систем при априорной неопределенности. Его значимость состоит в возможности создания универсальных математических средств, адаптируемых к условиям исследования систем различной природы [1 – 4].

Вычислительная эффективность непараметрических алгоритмов обработки информации во многом определяется объемом статистических данных и снижается по мере его увеличения, что затрудняет построение систем принятия решений в условиях больших выборок.

Естественным выходом в подобной ситуации является использование принципов декомпозиции исходных статистических данных по их объему и технологии параллельных вычислений. С этих позиций предложена и исследована смесь непараметрических оценок плотностей вероятности для одномерных и многомерных случайных величин [5 – 9]. Показано, что она имеет значительно меньшую дисперсию по сравнению с традиционной непараметрической оценкой

* Работа выполнена в рамках базовой части государственного задания Министерства образования и науки РФ высшим учебным заведениям на 2014 – 2016 гг. (СибГАУ № Б121/14).

плотности вероятности типа Розенблатта – Парзена. При этом сокращение времени вычислений сопоставимо с количеством составляющих смеси непараметрических оценок плотностей вероятности.

Полученные результаты обобщены при оценивании решающей функции в задаче распознавания образов для условий больших выборок. При этом разработаны двухуровневые непараметрические системы для решения задач классификации, установлены асимптотические свойства оценок их уравнений разделяющих поверхностей для одномерного и многомерного случаев [10 – 13].

Перспективное направление «обхода» проблем больших выборок связано с использованием регрессионной оценки плотности вероятности, синтез которой основан на декомпозиции исходных статистических данных и последующем анализе количественных характеристик получаемых множеств случайных величин [14, 15]. Оценка уравнения разделяющей поверхности, синтез которой осуществляется на основе статистик подобного типа, будем называть регрессионной.

Цель данной работы заключается в исследовании зависимости аппроксимационных свойств одномерной регрессионной оценки уравнения разделяющей поверхности в двуальтернативной задаче распознавания образов от методов дискретизации значений случайной величины в классах.

Регрессионная оценка плотности вероятности

Пусть имеется выборка $\overline{V} = \left(x^i, i = \overline{1, n}\right)$ из n независимых значений одномерной случайной величины x с неизвестной плотностью вероятности $p(x)$.

Разобьем область определения $p(x)$ на N непересекающихся интервалов длиной 2β и сформируем множества случайных величин X^j, $j = \overline{1, N}$. В качестве количественных характеристик X^j примем частоту $\overline{P}^j$ попадания случайной величины x в j-й интервал и его центр z^j. На основе полученной информации определим массив данных $\overline{V_1} = \left(z^j, y^j = \overline{P}^j/(2\beta), j = \overline{1, N}\right)$, составленный из центров z^j введенных интервалов и соответствующих им оценок y^j плотности вероятности. Объем N полученных данных может быть значительно меньше объема n исходной статистической информации.

В качестве приближения по эмпирическим данным $\overline{V_1}$ искомой плотности вероятности $p(x)$ примем статистику [14]

$$\overline{\varphi}(x) = c^{-1} \sum_{j=1}^{N} \overline{P}^j \Phi\left(\frac{x - z^j}{c}\right), \tag{1}$$

где ядерные функции $\Phi(u)$ удовлетворяют условиям H [16]:

$$\Phi(u) = \Phi(-u), \quad 0 \le \Phi(\mathrm{u}) < \infty, \ \int \Phi(u)\,du = 1, \quad \int u^2 \Phi(u)\,du = 1.$$

Здесь и далее бесконечные пределы опускаются.

Коэффициенты размытости $c = c(N)$ ядерных функций в оценке плотности вероятности (1) убывают с ростом количества N интервалов дискретизации области определения плотности вероятности $p(x)$.

Статистику (1) нетрудно получить, подставляя в выражение условного математического ожидания

$$\varphi(x) = \int y\, p\left(y \big/ x \right) dy \qquad (2)$$

непараметрическую оценку совместной плотности вероятности случайных величин (y, z)

$$\bar{p}(y, x) = \frac{1}{n c c_1} \sum_{j=1}^{N} \Phi\left(\frac{x - z^j}{c} \right) \Phi\left(\frac{y - y^j}{c_1} \right)$$

и известную плотность вероятности $(2 \beta N)^{-1}$ распределения центров z интервалов дискретизации.

Нетрудно убедиться, что регрессионная оценка плотности $\bar{\varphi}(x)$ является нормированной функцией, т.е. удовлетворяет основному свойству плотности вероятности. Регрессионная оценка плотности вероятности $\bar{\varphi}(x)$ обладает свойствами асимптотической сходимости к оптимальной модели $\varphi(x)$ (2) [14].

Синтез непараметрического алгоритма распознавания образов

Рассмотрим методику построения непараметрического классификатора, соответствующего критерию максимального правдоподобия, с использованием регрессионной оценки плотности вероятности (1).

Пусть $V = V_1 \bigcup V_2$ – обучающая выборка, составленная из значений признака x классифицируемых объектов $V_j = \left(x^i, i = \overline{1, n_j} \right)$, принадлежащих к одному из двух классов $\Omega_j = \Omega_j(x)$, $j = 1, 2$. Вид условных плотностей вероятностей $p_j(x)$ распределения значений x в классах Ω_j, $j = 1, 2$ неизвестен.

В данных условиях непараметрическое решающее правило распознавания образов, соответствующее критерию максимального правдоподобия, имеет вид

$$\bar{m}(x): \begin{cases} x \in \Omega_1 \,, \; \textit{если } \bar{f}_{12}(x) < 0 \\ x \in \Omega_2 \,, \; \textit{если } \bar{f}_{12}(x) \ge 0 \,, \end{cases} \qquad (3)$$

где $\bar{f}_{12}(x) = \bar{\varphi}_2(x) - \bar{\varphi}_1(x)$ – регрессионная оценка уравнения разделяющей поверхности

$$f_{12}(x) = \varphi_2(x) - \varphi_1(x) \qquad (4)$$

между классами Ω_1, Ω_2.

В качестве оценки условной плотности вероятности $p_j(x)$ будем использовать статистику типа (1). Для этого воспользуемся методикой декомпозиции области определения $p_j(x)$, представленной в предыдущем разделе. На этой основе преобразуем исходные данные V_j в выборку $\bar{V}_j = \left(z^i, \bar{P}_j^i \big/ (2\beta_j), i = \overline{1, N_j} \right)$. Здесь z^i – центр i-го интервала дискретизации области определения $p_j(x)$ длинной $2\beta_j$; N_j – их количество; $\bar{P}_j^i$ – частота попадания случайной величины x в i-й интервал.

Тогда непараметрическая оценка $\bar{f}_{12}(x)$, восстанавливаемая по выборкам $\overline{V}_j$, $j=1,2$, запишется в виде

$$\bar{f}_{12}(x)=c^{-1}\sum_{j=1}^{2}(-1)^j\sum_{i=1}^{N_j}\overline{P}_j^i\ \Phi\left(\frac{x-z^i}{c}\right). \qquad (5)$$

Оптимизация непараметрического решающего правила (3) по коэффициенту размытости c ядерных функций осуществляется в режиме «скользящего экзамена» из условия минимума статистической оценки вероятности ошибки распознавания образов

$$\overline{\rho}(c)=\frac{1}{N}\sum_{t=1}^{N}1\left(\sigma(t),\overline{\sigma}(t)\right),\ 1\left(\sigma(t),\overline{\sigma}(t)\right)=\begin{cases}0,\, если\ \ \sigma(t)=\overline{\sigma}(t),\\ \\ 1,\, если\ \ \sigma(t)\neq\overline{\sigma}(t),\end{cases}$$

где $N=N_1+N_2$; $\sigma(t),\overline{\sigma}(t)$ – соответственно «указания учителя» и решение алгоритма (3) о принадлежности ситуации z^t к одному из двух классов. При формировании решения $\overline{\sigma}(t)$ ситуация z^t исключается из процесса обучения в непараметрической статистике (5).

Асимптотические свойства непараметрической оценки уравнения разделяющей поверхности

При анализе асимптотических свойств статистики (5) будем использовать технологию преобразований, предложенную в работе [16] и развитую в статьях [17 – 20].

Во-первых. Вычислим математическое ожидание $\bar{f}_{12}(x)$. По определению имеем:

$$\begin{aligned} M\left(\bar{f}_{12}(x)\right)&=\frac{2\beta_2}{c}\sum_{i\in I_2}\iint y_2^i\Phi\left(\frac{x-z_2^i}{c}\right)p_2\left(y_2^i,z_2^i\right)dy_2^i\,dz_2^i-\\ &-\frac{2\beta_1}{c}\sum_{i\in I_1}\iint y_1^i\Phi\left(\frac{x-z_1^i}{c}\right)p_1\left(y_1^i,z_1^i\right)dy_1^i\,dz_1^i=\\ &=\frac{2}{c}\sum_{j=1}^{2}(-1)^j\beta_j\sum_{i\in I_j}\iint y_j^i\Phi\left(\frac{x-z_j^i}{c}\right)p_j\left(y_j^i,z_j^i\right)dy_j^i\,dz_j^i= \qquad (6)\\ &=\frac{1}{c}\sum_{j=1}^{2}(-1)^j\iint y_j\ p_j\left({y_j}/{t}\right)dy_j\ \Phi\left(\frac{x-t}{c}\right)p_j(t)\,dt=\\ &=\frac{1}{c}\sum_{j=1}^{2}(-1)^j\int\varphi_j(t)\ \Phi\left(\frac{x-t}{c}\right)dt\,, \end{aligned}$$

где M – знак математического ожидания; I_j – множество номеров элементов обучающей выборки, принадлежащих классу Ω_j; $p_j(t)=\left(2\beta_j N_j\right)^{-1}$ – плотность вероятности распределения центров интервалов дискретизации.

При выполнении преобразований учитывается, что элементы промежуточной статистической выборки $\overline{V}_j$ являются значениями одних и тех же случайных величин (y_j, t) с плотностью вероятности $p_j(y_j, t)$. Причем $\varphi_j(t) = \int y_j \, p\left(y_j / t \right) dy_j$ – кривая регрессии y_j по t, которая представляет собой оптимальное решающее правило в смысле минимума среднеквадратического отклонения при восстановлении неизвестной плотности вероятности $p_j(x)$.

Проведем в выражении (6) замену переменных $(x-t)c^{-1} = u$ и разложим функции $\varphi_j(x - cu)$ в ряд Тейлора в точке x. Тогда с учетом свойств H ядерных функций при достаточно больших значениях N_j, $j = \overline{1,2}$ получим асимптотическое выражение смещения регрессионной оценки решающей функции $\bar{f}_{12}(x)$ относительно $f_{12}(x)$ (4).

$$M\left(\bar{f}_{12}(x) - f_{12}(x)\right) \sim \frac{c^2}{2}\left(\varphi_2^{(2)}(x) - \varphi_1^{(2)}(x)\right), \qquad (7)$$

где $\varphi_j^{(2)}(x)$ – вторая производная функции $\varphi_j(x)$, $j = \overline{1,2}$ по x. Отсюда, из условия $c \to 0$ при $N_j \to \infty$, $j = \overline{1,2}$, следует свойство асимптотической несмещенности регрессионной оценки решающей функции $\bar{f}_{12}(x)$.

Во-вторых. Определим асимптотические свойства усредненной квадратической ошибки аппроксимации

$$M\int\left(\bar{f}_{12}(x) - f_{12}(x)\right)^2 dx = M\int\left(\varphi_2(x) - \overline{\varphi}_2(x)\right)^2 dx -$$
$$- 2M\int\left(\varphi_1(x) - \overline{\varphi}_1(x)\right)\left(\varphi_2(x) - \overline{\varphi}_2(x)\right)dx + M\int\left(\varphi_1(x) - \overline{\varphi}_1(x)\right)^2 dx. \qquad (8)$$

Асимптотические выражения для среднеквадратического отклонения $\overline{\varphi}_j(x)$ от $\varphi_j(x)$ получено в работе [14]

$$M\int\left(\varphi_j(x) - \overline{\varphi}_j(x)\right)dx \sim \frac{c^2}{2}\varphi_j^{(2)}(x), \qquad (9)$$

$$M\int\left(\varphi_j(x) - \overline{\varphi}_j(x)\right)^2 dx \sim \frac{2\Delta_j}{N_j c}\left\|\varphi_j(x)\right\|^2 \left\|\Phi(u)\right\|^2 + \frac{c^4}{4}\left\|\varphi_j^{(2)}(x)\right\|^2, \; j = \overline{1,2}, \qquad (10)$$

где $\left\|\varphi_j(x)\right\|^2 = \int\left(\varphi_j(x)\right)^2 dx$, $\left\|\Phi(u)\right\|^2 = \int \Phi^2(u)du$, $\left\|\varphi_j^{(2)}(x)\right\|^2 = \int\left(\varphi_j^{(2)}(x)\right)^2 dx$.

Тогда при достаточно больших значениях N_j, $j = \overline{1,2}$ асимптотическое выражение для среднеквадратического отклонения (8) запишется в виде

$$M\int\left(\bar{f}_{12}(x) - f_{12}(x)\right)^2 dx \sim \frac{2\left\|\Phi(u)\right\|^2}{c}\sum_{j=1}^{2}\frac{\Delta_j}{N_j}\left\|\varphi_j(x)\right\|^2 + \frac{c^4}{4}\left\|\varphi_2^{(2)}(x) - \varphi_1^{(2)}(x)\right\|^2. \; (11)$$

Нетрудно заметить, что при выполнении условий $c \to 0$, $N_j c \to \infty$ при

$N_j \to \infty$, $j = \overline{1,2}$ непараметрическая оценка $\bar{f}_{12}(x)$ сходится в среднеквадратическом к уравнению разделяющей поверхности (4), а с учетом свойства ее асимптотической несмещенности является состоятельной оценкой.

Анализ асимптотических свойств статистики $\bar{f}_{12}(x)$

Определим минимальное значение W_2 выражения (11) при оптимальном значении $\bar{c}$ коэффициентов размытости ядерных функций непараметрической оценки уравнения разделяющей поверхности $\bar{f}_{12}(x)$. Приравнивая производную выражения (11) по c к нулю, получим:

$$\bar{c} = \left(\frac{2\,\|\Phi(u)\|^2 \left(\Delta_1\, N_2\, \|\varphi_1(x)\|^2 + \Delta_2\, N_1\, \|\varphi_2(x)\|^2\right)}{N_1\, N_2\, \left\|\varphi_2^{(2)}(x) - \varphi_1^{(2)}(x)\right\|^2} \right)^{1/5} .$$

Тогда при $c = \bar{c}$ выражение (11) перепишется в виде

$$W_2 = \frac{5}{4}\left[\left(2\,\|\Phi(u)\|^2 \sum_{j=1}^{2} \frac{\Delta_j}{N_j} \|\varphi_j(x)\|^2\right)^4 \left\|\varphi_2^{(2)}(x) - \varphi_1^{(2)}(x)\right\|^2\right]^{1/5} . \tag{12}$$

Исследуем зависимость аппроксимационных свойств регрессионной оценки уравнения разделяющей поверхности (5) от методов дискретизации значений x в классах. Для этого определим отношение асимптотических выражений среднеквадратических отклонений (12), в которых количество N_j, $j = \overline{1,2}$ интервалов дискретизации вычисляются в соответствии с методикой

$$N_j = \sqrt{\Delta_j \left\|p_j(x)\right\|^2}\,\sqrt{n_j}\ , \tag{13}$$

предложенной в работах [21, 22], и формулами Хайнкольда и Гаеде

$$N_j = \sqrt{n_j}\ , \tag{14}$$

Брукса и Каррузера

$$N_j = 5\lg n_j , \tag{15}$$

Старджесса

$$N_j = \log_2 n_j + 1,\ j = \overline{1,2}. \tag{16}$$

Будем считать, что количество элементов обучающей выборки, принадлежащих классам, одинаковы, т.е. $n_1 = n_2 = \bar{n}$. Причем плотности вероятности $p_j(x)$ признака x в классах имеют нормальный закон распределения. В этом случае значение $\sqrt{\Delta_j \left\|p_j(x)\right\|^2} = 1.3$, $j = \overline{1,2}$.

Обозначим через W_2 (13) значение асимптотического выражения среднеквадратического критерия (12), полученное, например, при использовании формулы дискретизации (13).

Тогда нетрудно показать, что

$$R(13,14)=\frac{W_2(13)}{W_2(14)}=0.9,$$

$$R(13,15)=\frac{W_2(13)}{W_2(15)}=\left(\frac{5\lg\bar{n}}{1.3\sqrt{\bar{n}}}\right)^{\frac{4}{5}},$$

$$R(13,16)=\frac{W_2(13)}{W_2(16)}=\left(\frac{\log_2\bar{n}+1}{1.3\sqrt{\bar{n}}}\right)^{\frac{4}{5}}.$$

Анализ полученных отношений свидетельствует, что при использовании метода дискретизации (13) достигаются более высокие аппроксимационные свойства регрессионной оценки уравнения разделяющей поверхности по сравнению с применением формул (14) – (16) (рис. 1). Данный вывод согласуется с результатами исследований работы [21], в которой обосновывается оптимальный выбор количества интервалов дискретизации области определения плотности вероятности.

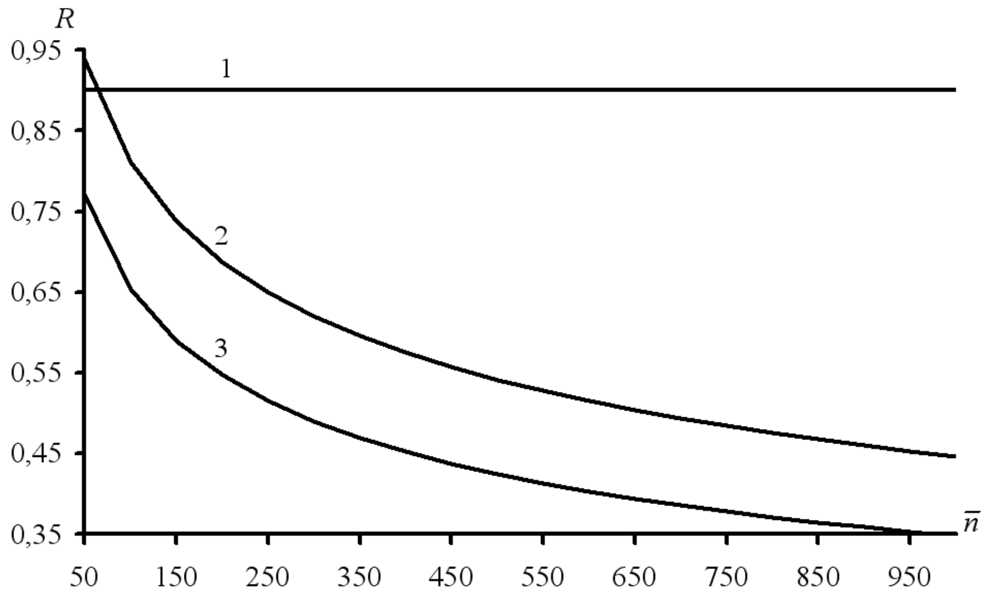

Рис. 1. Зависимости значений R от количества $\bar{n}$ элементов обучающей выборки в условиях $n_1=n_2=\bar{n}$, где кривые 1, 2, 3 соответствуют отношениям $R(13,14)$, $R(13,15)$, $R(13,16)$.

С ростом объема обучающей выборки преимущество метода дискретизации (13) над рекомендациями (15), (16) возрастает.

Заключение

Регрессионная оценка уравнения разделяющей поверхности является основой построения непараметрических алгоритмов распознавания образов в условиях обучающих выборок большого объема. Ее синтез осуществляется путем декомпозиции исходной информации и анализа количественных характеристик, получаемых множеств случайных величин на основе кривой регрессии. Предлагаемая статистика обладает свойством асимптотической несмещенности и состоятельности относительно оптимального решающего правила. Методика оптимального выбора количества интервалов дискретизации области определения плотности вероятности является эффективной и при оценивании уравнения разделяющей поверхности между классами в задаче распознавания образов. К данному методу по своим показателем близка формула Хайнкольда и Гаеде. Применение формул Брукса и Каррузера, а также формулы Старджесса сопряжено с ухудшением аппроксимационных свойств регрессионной оценки уравнения разделяющей поверхности, что особенно отчетливо проявляется с увеличением объема обучающей выборки.

Получаемые результаты открывают возможность их распространения на проблему синтеза и анализа многомерной регрессионной оценки уравнения разделяющей поверхности.

ЛИТЕРАТУРА

1. *Лапко А.В., Лапко В.А.* Непараметрические системы обработки неоднородной информации. – Новосибирск: Наука, 2007.
2. *Лапко А.В., Лапко В.А.* Анализ непараметрических алгоритмов распознавания образов в условиях пропуска данных // Автометрия. – 2008. – Т. 44, № 3. – С. 65-74.
3. *Лапко А.В., Лапко В.А.* Применение непараметрического алгоритма распознавания образов в задаче проверки гипотезы о распределениях случайных величин // Системы управления и информационные технологии. – 2010. – Т. 41, №3. – С. 8-11.
4. *Лапко В.А., Капустин А.Н.* Синтез нелинейных непараметрических коллективов решающих правил в задачах распознавания образов // Автометрия. – 2006. – Т.42, №6. – С.26-33.
5. *Лапко А.В., Лапко В.А., Егорочкин И.А.* Непараметрические оценки смеси плотностей вероятности и их применение в задаче распознавания образов // Системы управления и информационные технологии. – 2009. – Т. 35, № 1. – С. 60-64.
6. *Лапко А.В., Лапко В.А.* Анализ свойств смеси непараметрических оценок плотности вероятности многомерной случайной величины // Вестник СибГАУ. 2010. – Т.28, №2. – С. 32-35.
7. *Лапко А. В., Лапко В. А.* Свойства непараметрической оценки плотности вероятности многомерных случайных величин в условиях больших выборок // Информатика и системы управления. – 2012. – Т.32, №2. – С. 121-126.
8. *Лапко А.В., Лапко В.А.* Синтез структуры смеси непараметрических оценок плотности вероятности многомерной случайной величины // Системы управления и информационные технологии. – 2011. – Т.43, №1. – С. 12-15.
9. *Лапко А.В., Лапко В.А.* Анализ свойств непараметрических оценок смеси плотностей вероятности при различных условиях распределения статистических данных // Информатика и системы управления. – 2013. – Т.35, №1. – С. 119-126.
10. *Лапко А.В., Лапко В.А.* Разработка и исследование двухуровневых непараметрических систем классификации // Автометрия. – 2010. – Т. 46, № 1. – С. 70-78.
11. *Лапко А.В., Лапко В.А.* Синтез структуры семейства непараметрических решающих функций в задаче распознавания образов // Автометрия. – 2011. – Т. 47, № 4. – С. 76-82.
12. *Лапко А.В., Лапко В.А.* Коллектив непараметрических решающих функций в двуальтернативной задаче распознавания образов // Системы управления и информационные технологии. – 2009. – Т. 37, №3.1. – С. 156 – 160.
13. *Лапко А.В., Лапко В.А.* Непараметрическая оценка уравнения разделяющей поверхности в условиях больших выборок и ее свойства // Системы управления и информационные технологии. – 2010. – Т. 39, № 1.2. – С. 300-304.
14. *Лапко А.В., Лапко В.А.* Регрессионная оценка плотности вероятности и ее свойства // Системы управления и информационные технологии. – 2012. – Т. 49, №3.1. – С. 152-156.

15. *Лапко А.В., Лапко В.А.* Непараметрические методики анализа множеств случайных величин // Автометрия. – 2003. – Т. 39, №1. – С.54-61.
16. *Епанечников В.А.* Непараметрическая оценка многомерной плотности вероятности // Теория вероятности и ее применения. – 1969. – Т.14, №1. – С. 156-161.
17. *Лапко А.В., Лапко В.А.* Анализ асимптотических свойств непараметрической оценки уравнения разделяющей поверхности в двуальтернативной задаче распознавания образов // Автометрия. – 2010. – Т. 46, № 3. – С. 48-53.
18. *Лапко А.В., Лапко В.А.* Асимптотические свойства многомерной непараметрической оценки уравнения разделяющей поверхности в двуальтернативной задаче распознавания образов // Системы управления и информационные технологии. – 2010. – Т.39, №1. – С. 16-19.
19. *Лапко А.В., Лапко В.А.* Асимптотические свойства непараметрической оценки уравнения разделяющей поверхности в алгоритме распознавания образов, соответствующего критерию максимума апостериорной вероятности // Системы управления и информационные технологии. – 2010. – №4 (42). – С. 58-61.
20. *Лапко А.В., Лапко В.А.* Непараметрическая оценка двумерного уравнения разделяющей поверхности в условиях пропуска данных и ее свойства // Информатика и системы управления. – 2013. – Т.36, №2. – С. 117-126.
21. *Лапко А.В., Лапко В.А.* Оптимальный выбор количества интервалов дискретизации области изменения одномерной случайной величины при оценивании плотности вероятности // Измерительная техника. – 2013. – №7. – С. 24 – 27.
22. *Lapko A.V., Lapko V.A.* Optimal selection of the number of sampling intervals in domain of variation of a one-dimensional random variable in estimation of the probability density // Measurement Techniques. – 2013. – Vol. 56, No. 7. – С. 24-27. (DOI: 10.1007/s11018-013-0279-x).

E-mail:
Лапко Александр Васильевич – lapko@icm.krasn.ru;
Лапко Василий Александрович – lapko@icm.krasn.ru.