УДК 681.513

А. В. Лапко, В. А. Лапко

НЕПАРАМЕТРИЧЕСКИЕ АЛГОРИТМЫ РАСПОЗНАВАНИЯ ОБРАЗОВ В ЗАДАЧЕ ПРОВЕРКИ ГИПОТЕЗЫ О РАСПРЕДЕЛЕНИЯХ СЛУЧАЙНЫХ ВЕЛИЧИН

Предлагается методика проверки гипотез о тождественности законов распределения случайных величин, основанная на использовании непараметрических алгоритмов распознавания образов и принципов коллективного оценивания. Приводятся результаты сравнения методики с критерием Колмогорова — Смирнова.

Ключевые слова: *непараметрическая статистика, распознавание образов, проверка гипотез, распределения случайных величин.*

Проверка гипотез о распределениях случайных величин является одной ключевых задач математической статистики и имеет важное прикладное значение, например, при сравнении эффективности приборов и систем контроля с данными их испытаний. Для проверки гипотез о распределениях случайных величин широко используется критерий согласия Пирсона, который не зависит от распределений случайных величин и их размерности [1]. Однако методика формирования критерия Пирсона содержит трудноформализуемый этап разбиения области возможных значений случайной величины на многомерные интервалы. Данный этап не отражен в критерии Колмогорова — Смирнова, который позволяет проверять гипотезы о распределениях одномерных случайных величин [2].

В работе [3] показана возможность использования непараметрических алгоритмов распознавания образов, соответствующих критерию максимального правдоподобия,

в задаче проверки статистических гипотез о распределениях случайных величин. Результаты использования предлагаемой методики сопоставимы с критерием Колмогорова — Смирнова для одномерных задач в условиях, когда число элементов сравниваемых последовательностей случайных величин различается незначительно. При неравных объемах случайных последовательностей эффективность предлагаемой методики снижается. Данный факт согласуется с результатами исследований [4], где показано значительное ухудшение аппроксимационных свойств непараметрической оценки уравнения разделяющей поверхности между классами при увеличении степени неравномерности распределения элементов обучающей выборки.

Цель исследования, описываемого в настоящей статье, — развитие данной методики на основе использования принципов коллективного оценивания при синтезе непараметрических алгоритмов распознавания образов.

Модифицированная методика проверки гипотезы о распределениях случайных величин. Пусть X_1 и X_2 — генеральные совокупности с произвольными законами распределения. Необходимо по независимым выборкам $V_1 = x^i, i = \overline{1, n_1}$, и $V_2 = x^i, i = \overline{1, n_2}$, полученным из данных генеральных совокупностей, проверить либо опровергнуть гипотезу

$$H_\circ : P_1(x) \equiv P_2(x)$$

о тождественности функций распределения.

Известно, что если при решении двухальтернативной задачи распознавания образов вероятность ошибки классификации равна 0,5, то законы распределения случайных величин в области определения классов совпадают. Поэтому появляется возможность перехода от задачи сравнения законов распределения случайных величин к проверке гипотезы $\bar{H}_\circ$ о равенстве статистической оценки вероятности ошибки распознавания образов значению 0,5.

При реализации предлагаемой модифицированной методики необходимо выполнить следующие действия.

1. Пусть число элементов сравниваемых последовательностей случайных величин отличается значительно, например $n_1 > n_2$. Требуется сформировать совокупность сравниваемых последовательностей $V_1(j) = x^i, i \in I_j, V_2 = x^i, i = \overline{1, n_2}$, $j = \overline{1, T}$. Элементы выборки $V_1(j)$ объемом n_2 формируются случайным образом из последовательности V_1; здесь I_j — множество номеров элементов последовательности V_1, составляющих последовательность $V_1(j)$. Присвоим элементам множества I_j значения $n_2 + t, t = \overline{1, n_2}$.

2. На основе множеств $V_1(j), V_2$ определить обучающую выборку $V(j) = \left(x^i, \sigma(i), i = \overline{1, 2n_2}\right)$ для решения задачи распознавания образов, где

$$\sigma(i) = \begin{cases} -1 \; \forall \; x^i \in \Omega_1, \\ 1 \; \forall \; x^i \in \Omega_2 \end{cases}$$

свидетельствует о принадлежности значения x^i к тому либо иному классу Ω_1, Ω_2. При этом полагаем, что элементы множеств $V_1(j)$ и V_2 принадлежат соответственно классам Ω_1, Ω_2.

3. По выборке $V(j)$ осуществить синтез непараметрического алгоритма распознавания образов, соответствующего критерию максимального правдоподобия [5]:

$$\overline{m}_j(x):\begin{cases} x\in\Omega_1 \quad \forall \quad \overline{f}_{12}^{\,j}(x)\le 0, \\ x\in\Omega_2 \quad \forall \quad \overline{f}_{12}^{\,j}(x)>0. \end{cases} \tag{1}$$

При формировании оценки уравнения разделяющей поверхности

$$\overline{f}_{12}^{\,j}(x)=\overline{p}_2(x)-\overline{p}_1^{\,j}(x) \tag{2}$$

будем использовать непараметрические оценки

$$\overline{p}_2(x)=(n_2\ c)^{-1}\sum_{i=1}^{n_2}\Phi\left(\frac{x-x^i}{c}\right),$$

$$\overline{p}_1^{\,j}(x)=(n_2\ c)^{-1}\sum_{i=n_2+1}^{2n_2}\Phi\left(\frac{x-x^i}{c}\right)$$

плотностей вероятности распределения x в классах Ω_1, Ω_2 типа Розенблатта — Парзена [6]. Ядерные функции $\Phi(u)$ удовлетворяют условиям $\Phi(u)=\Phi(-u)$, $0\le\Phi(u)<\infty$, $\int_{-\infty}^{+\infty}\Phi(u)du=1$, а значения их коэффициентов размытости c убывают с увеличением n_2.

Тогда статистика (2) может быть представлена выражением

$$\tilde{f}_{12}^{\,j}(x)=(n_2\ c)^{-1}\sum_{i=1}^{2n_2}\sigma(i)\Phi\left(\frac{x-x^i}{c}\right). \tag{3}$$

Выбор оптимального значения $\overline{c}$ коэффициента размытости непараметрического решающего правила $\overline{m}_j(x)$ осуществляется согласно условию минимума оценки вероятности ошибки распознавания образов

$$\overline{\rho}_j(c)=\frac{1}{2n_2}\sum_{t=1}^{2n_2}1\left(\sigma(t),\overline{\sigma}(t)\right),$$

где индикаторная функция

$$1\left(\sigma(t),\overline{\sigma}(t)\right)=\begin{cases} 0 \quad \forall \quad \sigma(t)=\overline{\sigma}(t); \\ 1 \quad \forall \quad \sigma(t)\ne\overline{\sigma}(t), \end{cases}$$

здесь $\overline{\sigma}(t)$ — „решение“ алгоритма $\overline{m}_j(x)$ о принадлежности значений x^t к тому либо иному классу Ω_1, Ω_2, полученное в соответствии с правилом (1).

При вычислении $\overline{\rho}_j(c)$ „решение“ $\overline{\sigma}(t)$ алгоритма (1) определяется в соответствии со знаком статистики

$$\tilde{f}_{12}^{\,j}(x^t)=(n_2\ c)^{-1}\sum_{\substack{i=1\\ i\ne t}}^{2n_2}\sigma(i)\Phi\left(\frac{x^t-x^i}{c}\right),$$

т.е. значение x^t исключается.

4. Проверить гипотезу $\overline{H}_\circ(j):\ \overline{\rho}_j(\overline{c})=0{,}5$ в соответствии с критерием Колмогорова — Смирнова. Для этого сравним его пороговое значение [7]

$$D_\alpha = \sqrt{-\ln\left(\frac{\alpha}{2}\right)\cdot\frac{1}{4n_2}}$$

с отклонением $\bar{D}_{12}^{j} = \left|0,5 - \bar{\rho}_j\left(\bar{c}\right)\right|$; здесь α — вероятность (риск) отвергнуть правильную гипотезу $\bar{H}_\circ\left(j\right)$.

Если выполняется соотношение $\bar{D}_{12}^{j} < D_\alpha$, то гипотеза $\bar{H}_\circ\left(j\right)$ справедлива, иначе — она отвергается.

5. В соответствии с пп. 2—4 проверить гипотезы $\bar{H}_\circ\left(j\right)$ на основе последовательностей случайных величин $V_1\left(j\right), V_2$, $j = \overline{1,T}$. По полученным данным рассчитать оценки вероятностей $\bar{P}_1 = S/T$, $\bar{P} = \bar{S}/T$ справедливости гипотезы $\bar{H}_\circ$ и ее отклонения соответственно. Здесь S — количество „решений“ о справедливости гипотез $\bar{H}_\circ\left(j\right)$, $j = \overline{1,T}$, а $\bar{S}$ — количество решений об их отклонении.

6. Проверить достоверность отличия оценок $\bar{P}_1$ и $\bar{P}$ с использованием критерия Колмогорова — Смирнова. Для этого вычислим его пороговое значение

$$D_\alpha = \sqrt{-\ln\frac{\alpha}{2}\Big/T},$$

которое сравним с разностью $\bar{D} = \left|\bar{P}\left(T\right) - \bar{P}_1\left(T\right)\right|$.

Исходная гипотеза $H_\circ$ подтверждается, если $\bar{D} > D_\alpha$ и $\bar{P}_1 > \bar{P}$, в противном случае при $\bar{P}_1 < \bar{P}$ она отвергается.

Анализ результатов экспериментов. Было проведено сравнение эффективности базовой [3] и модифицированной методик проверки гипотезы о распределениях случайных величин и критерия Колмогорова — Смирнова по данным вычислительных экспериментов. Последовательности $V_1 = x^i, i = \overline{1, n_1},$ и $V_2 = x^i, i = \overline{1, n_2},$ случайных наблюдений формировались на основе датчиков случайных величин с равномерным $x^i = \varepsilon^i$ и нормальным $x^i = 0,5 + 0,15\left(\sum_{j=1}^{12}\varepsilon^j - 6\right)$, $i = \overline{1,n}$, законами распределения. Случайные величины ε с равномерным законом распределения определены на интервале $\left[0,1\right]$. При их формировании использовался стандартный датчик псевдослучайных величин среды визуального программирования „Delphi“.

При фиксированных условиях исследования было проведено 100 вычислительных экспериментов. По полученным результатам при априори тождественных законах распределения случайных величин оценивалась вероятность P_0 справедливости гипотезы $H_\circ$. Если законы распределения отличались, оценивалась вероятность P_1 отклонения гипотезы $H_\circ$. Риск α отвергнуть гипотезу $H_\circ$ принимался равным 0,05.

При синтезе непараметрического классификатора использовались параболические ядерные функции Епанечникова [8].

Результаты вычислительного эксперимента при различных условиях проверки гипотезы о распределениях представлены на рис. 1 и 2: рис. 1 — зависимости оценок вероятностей P_0 справедливости гипотезы $H_\circ$ от объема экспериментальных данных $n = n_1 + n_2$

для $n_1 = 1,2n_2$ (*а*) и $n_1 = 2n_2$ (*б*) при сравнении двух априори тождественных нормальных законов распределения случайных величин; кривая *1* получена при использовании критерия Колмогорова — Смирнова, кривая *2* — базовой методики [3], кривая *3* — модифицированной методики при $T = 10$; рис. 2 — зависимости оценок вероятностей P_1 отклонения гипотезы $H_\circ$ от объема экспериментальных данных $n = n_1 + n_2$ для $n_1 = 2n_2$ при сравнении равномерного и нормального законов распределения (обозначения кривых соответствуют принятым для рис. 1).

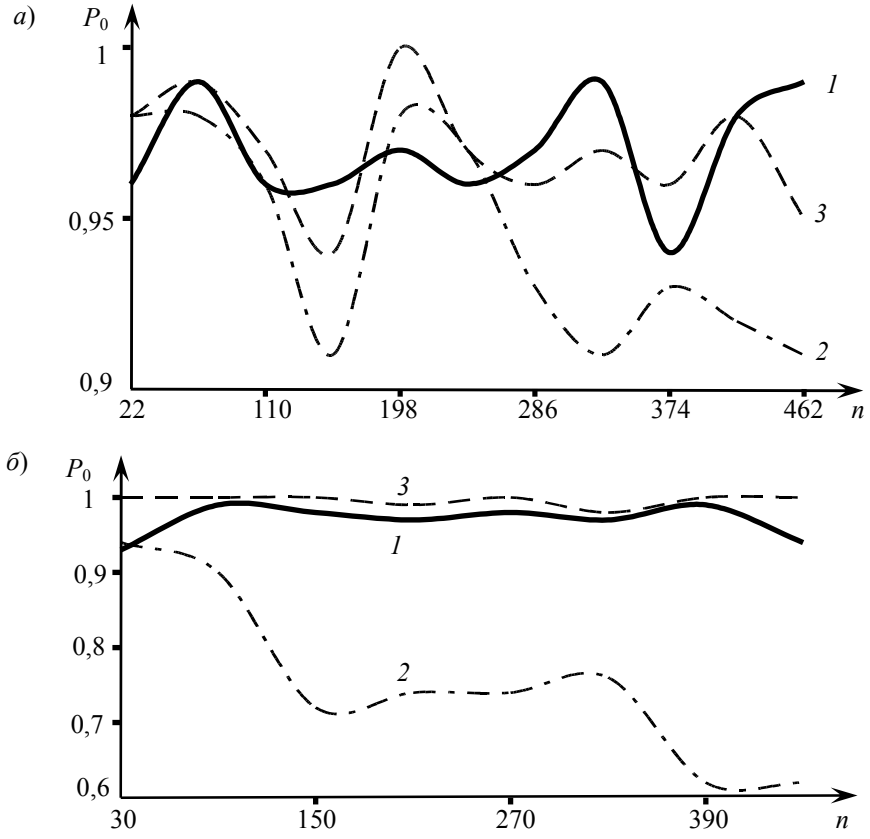

Рис. 1

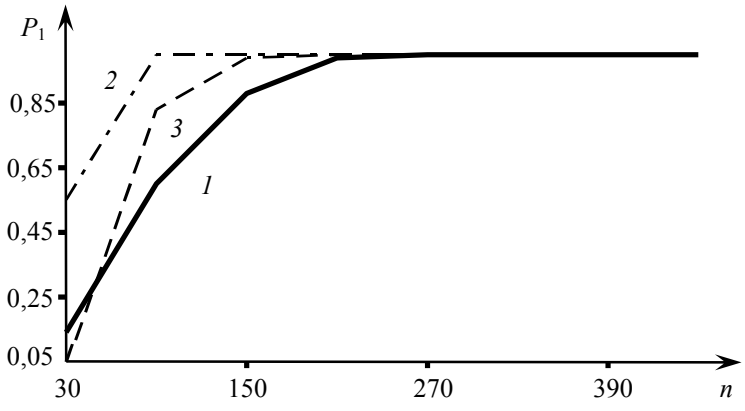

Рис. 2

Представленные графики подтверждают эффективность модифицированной методики.

Заключение. Применение рассмотренной модифицированной методики позволяет расширить условия проверки гипотез о распределениях. Эффективность предложенной методики сопоставима с критерием Колмогорова — Смирнова для одномерных задач. Полученные результаты могут быть обобщены для задачи проверки гипотез о тождественности законов распределения многомерных случайных величин.

Исследования, результаты которых представлены в настоящей статье, выполнены в рамках Федеральной целевой программы „Научные и научно-педагогические кадры инновационной России“ на 2009—2013 гг., гос. контракт № 02.740.11.0621.

СПИСОК ЛИТЕРАТУРЫ

1. *Пугачев В. С.* Теория вероятностей и математическая статистика. М.: Наука, 1979.

2. *Смирнов Н. В.* Оценка расхождения между кривыми распределения в двух независимых выборках // Бюл. Моск. ун-та. 1930. Т. 2, № 2. С. 3—14.

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. № 3(41). С. 8—11.

4. *Лапко А. В., Лапко В. А.* Анализ асимптотических свойств непараметрической оценки уравнения разделяющей поверхности в двухальтернативной задаче распознавания образов // Автометрия. 2010. Т. 46, № 3. С. 48—53.

5. *Лапко А. В., Лапко В. А., Соколов М. И., Ченцов С. В.* Непараметрические системы классификации. Новосибирск: Наука, 2000.

6. *Parzen E.* On estimation of a probability density function and mode // Ann. Math. Statistic. 1962. Vol. 33, N 3. P. 1065—1076.

7. *Шаракшанэ А. С., Железнов И. Г., Ивницкий В. А.* Сложные системы. М.: Высш. школа, 1977.

8. *Епанечников В. А.* Непараметрическая оценка многомерной плотности вероятности // Теория вероятности и ее применения. 1969. Т. 14, вып. 1. С. 156—161.

Сведения об авторах

Александр Васильевич Лапко — д-р техн. наук, профессор; Институт вычислительного моделирования СО РАН, Красноярск; E-mail: lapko@icm.krasn.ru

Василий Александрович Лапко — д-р техн. наук, профессор; Сибирский государственный аэрокосмический университет им. акад. М. Ф. Решетнёва, кафедра космических средств и технологий, Красноярск; E-mail: lapko@icm.krasn.ru

Рекомендована СибГАУ

Поступила в редакцию
19.11.10 г.