

СРАВНЕНИЕ НЕПАРАМЕТРИЧЕСКИХ КРИТЕРИЕВ ПРОВЕРКИ ГИПОТЕЗ О РАСПРЕДЕЛЕНИЯХ СЛУЧАЙНЫХ ВЕЛИЧИН*

С позиций принципов коллективного оценивания развивается методика проверки гипотезы о тождественности законов распределений случайных величин, основанная на непараметрических алгоритмах распознавания образов. Проведено ее сравнение с критериями Смирнова и Пирсона.

Ключевые слова: непараметрическая статистика, распознавание образов, проверка статистических гипотез, распределение случайных величин, критерий Смирнова, критерий Пирсона.

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 трудно формализуемый этап разбиения области возможных значений случайной величины на многомерные интервалы. Данный этап отсутствует в критерии Смирнова [2] и методике, основанной на использовании непараметрических алгоритмов распознавания образов [3; 4].

Проведем сравнение эффективности указанных критериев при проверке гипотезы о тождественности законов распределения одномерных случайных величин по данным вычислительных экспериментов.

Традиционные непараметрические критерии. Пусть X_1 и X_2 – две генеральные совокупности с произвольными законами распределения.

Необходимо по независимым выборкам $V_1 = (x^i, i = \overline{1, n_1})$ и $V_2 = (x^i, i = \overline{1, n_2})$, извлеченным из данных генеральных совокупностей, проверить либо опровергнуть гипотезу

$$H_0: P(X_1) \equiv P(X_2)$$

о тождественности законов распределения.

Методика проверки статистической гипотезы H_0 на основе критерия Смирнова сводится к выполнению следующих шагов.

Шаг 1. По независимым выборкам V_1, V_2 построить оценки функций распределения

$$\bar{P}_j(x) = \frac{1}{n_j} \sum_{i=1}^{n_j} 1(x - x^i), \quad j = 1, 2,$$

где

$$1(x - x^i) = \begin{cases} 0, & \text{если } x - x^i < 0, \\ 1, & \text{если } x - x^i \geq 0. \end{cases}$$

Шаг 2. Найти максимальное расхождение между эмпирическими функциями распределения

$$\bar{D}_{12} = \max_x |\bar{P}_1(x) - \bar{P}_2(x)|.$$

Шаг 3. В соответствии с критерием Смирнова [5] сравнить полученное максимальное расхождение $\bar{D}_{12}$ с пороговым:

$$D_\alpha = \sqrt{-\ln \frac{\alpha}{2} \left(\frac{1}{n_1} + \frac{1}{n_2} \right) / 2},$$

где α – принятый уровень доверия, т. е. риск отвергнуть гипотезу H_0 .

Если выполняется условие $\bar{D}_{12} < D_\alpha$, то гипотеза H_0 справедлива, иначе эмпирические законы распределения различаются.

Критерий Пирсона предполагает выполнение следующих шагов.

Шаг 1. Разбить область изменения исследуемых случайных величин на N непересекающихся интервалов. Их количество может быть определено по эвристическим формулам Старджесса

$$N = \log_2 n + 1,$$

Брукса и Каррузера

$$N = 5 \lg n$$

или Гаеде

$$N = \sqrt{n},$$

где $n = n_1 + n_2$.

Шаг 2. Вычислить частоты $\bar{P}_1^j, \bar{P}_2^j, \bar{P}_{12}^j$ попадания элементов последовательностей V_1, V_2 и $V_1 \cup V_2$ в каждый j -й интервал, $j = \overline{1, N}$. Если для некоторого j -го интервала значение $\bar{P}_{12}^j = 0$, то количество интервалов уменьшается на единицу и производится перерасчет размера интервалов и соответствующих им частот. Эта процедура повторяется до тех пор, пока не будет определено количество интервалов $\tilde{N} \leq N$, для которых выполняются условия $\bar{P}_{12}^j \neq 0, j = \overline{1, \tilde{N}}$.

*Работа выполнена при частичной поддержке гранта Федеральной целевой программы «Научные и научно-педагогические кадры инновационной России на 2009–2013 годы» (государственный контракт № 02.740.11.0621).

Шаг 3. Рассчитать значение случайной величины по формуле [1, с. 330]

$$Z = \sum_{i=1}^2 n_i \sum_{j=1}^{\tilde{N}} \frac{(\bar{P}_i^j - \bar{P}_{12}^j)^2}{\bar{P}_{12}^j},$$

которое имеет χ^2 -распределение с $k = \tilde{N} - 1$ степенями свободы.

Шаг 4. По таблице χ^2 -распределения определить порог $\chi^2(k, \alpha)$ одностороннего критерия Пирсона при заданных значениях k и уровне значимости α .

Гипотеза H_0 справедлива, если $Z < \chi^2(k, \alpha)$, иначе она отвергается.

Методика 1 проверки гипотезы о распределениях. Будем полагать, что элементы последовательностей случайных величин V_1 и V_2 принадлежат соответственно классам Ω_1 и Ω_2 . Сформируем обучающую выборку $V = (x^i, \sigma(i), i = \overline{1, n})$ для решения задачи распознавания образов, где $n = n_1 + n_2$; $\sigma(i)$ – указание о принадлежности значения x^i к тому либо иному классу. На этой основе построим непараметрический алгоритм распознавания образов, соответствующий критерию максимального правдоподобия [6]:

$$\bar{m}(x) : \begin{cases} x \in \Omega_1, & \text{если } \bar{f}_{12}(x) \leq 0, \\ x \in \Omega_2, & \text{если } \bar{f}_{12}(x) > 0. \end{cases} \quad (1)$$

При формировании оценки уравнения разделяющей поверхности

$$\bar{f}_{12}(x) = \bar{p}_2(x) - \bar{p}_1(x) \quad (2)$$

можно использовать непараметрические оценки $\bar{p}_1(x)$, $\bar{p}_2(x)$ плотностей вероятности распределения x в классах Ω_1 , Ω_2 типа Розенблатта–Парзена [7]. Тогда статистика (2) будет представлена выражением

$$\bar{f}_{12}(x) = (nc)^{-1} \sum_{i=1}^n \sigma_1(i) \Phi\left(\frac{x - x^i}{c}\right), \quad (3)$$

где

$$\sigma_1(i) = \begin{cases} -\bar{P}_1^{-1} \vee x^i \in \Omega_1, \\ \bar{P}_2^{-1} \vee x^i \in \Omega_2; \end{cases}$$

$\bar{P}_j = \frac{n_j}{n}$ – оценка априорной вероятности принадлежности ситуаций обучающей выборки к классу Ω_j , $j = 1, 2$. Ядерные функции в статистике (3) удовлетворяют условиям $\Phi(u) = \Phi(-u)$, $0 \leq \Phi(u) < \infty$, $\int_{-\infty}^{+\infty} \Phi(u) du = 1$, а значения их коэффициентов размытости c убывают с ростом количества элементов множеств V_j , $j = 1, 2$.

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

$$\bar{p}(c) = \frac{1}{n} \sum_{i=1}^n 1(\sigma(i), \bar{\sigma}(i)),$$

где индикаторная функция

$$1(\sigma(i), \bar{\sigma}(i)) = \begin{cases} 0 & \forall \sigma(i) = \bar{\sigma}(i) \\ 1 & \forall \sigma(i) \neq \bar{\sigma}(i); \end{cases}$$

$\bar{\sigma}(i)$ – решение о принадлежности значений x^i к классу Ω_1 либо Ω_2 , полученное в соответствии с алгоритмом распознавания образов (1).

При вычислении $\bar{p}(c)$ решение $\bar{\sigma}(i)$ алгоритма (1) определим в соответствии со знаком статистики

$$\tilde{f}_{12}(x^i) = (nc)^{-1} \sum_{\substack{i=1 \\ i \neq t}}^n \sigma_1(i) \Phi\left(\frac{x^i - x^t}{c}\right),$$

т. е. ситуация x^t , которая подается на контроль, исключается из процесса обучения.

Известно, что если при решении двувальтернативной 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 $\bar{H}_0$ о равенстве статистической оценки вероятности ошибки распознавания образов значению 0,5.

Проверим гипотезу $\bar{H}_0 : \bar{p}(\bar{c}) = 0,5$ в соответствии с критерием Колмогорова. Для этого сравним его пороговое значение

$$\tilde{D}_\alpha = \sqrt{-\ln \frac{\alpha}{2} \left(\frac{1}{n_1 + n_2} \right) / 2}$$

с отклонением $\bar{D}_{12} = |0,5 - \bar{p}(\bar{c})|$ при вероятности α отвергнуть правильную гипотезу $\bar{H}_0$.

Гипотеза $\bar{H}_0$ справедлива при выполнении условия $\bar{D}_{12} < \tilde{D}_\alpha$, иначе она отвергается.

Существуют условия, когда использование предлагаемой методики и критерия Смирнова приводит к сопоставимым результатам. К таким условиям относятся задачи проверки гипотез при разных законах распределения случайных величин и одинаковых законах распределения, когда объемы сравниваемых последовательностей V_1 , V_2 отличаются незначительно [3].

Для различных объемах случайных последовательностей, например при $n_1 = 2n_2$, установлено снижение эффективности методики 1 по сравнению

с критерием Смирнова. Данный факт согласуется с результатами работы [8], где показано значительное снижение аппроксимационных свойств непараметрической оценки уравнения разделяющей поверхности при увеличении степени неравномерности распределения элементов обучающей выборки между классами.

Методика 2 проверки гипотезы о распределениях. Пусть количество элементов исследуемых последовательностей случайных величин отличается значительно, например $n_1 > n_2$. Сформируем набор сравниваемых последовательностей $(V_1(j) = (x^i, i \in I_j), V_2 = (x^i, i = \overline{1, n_2}))$, $j = \overline{1, T}$, где I_j – множество номеров элементов последовательности V_1 , составляющих сравниваемую последовательность $V_1(j)$. Элементы выборки $V_1(j)$ объемом n_2 формируются случайным образом из последовательности V_1 .

В соответствии с методикой 1 проверим гипотезы $\bar{H}_0(j)$: $\bar{p}_j(\bar{c}) = 0,5$ и по полученным данным рассчитаем оценки вероятностей $\bar{P}_0 = S/T$, $\bar{P} = \bar{S}/T$ справедливости гипотезы $\bar{H}_0$ и ее отклонения соответственно. Здесь S – количество решений о справедливости, а $\bar{S}$ – об отклонении гипотез $\bar{H}_0(j)$, $j = \overline{1, T}$.

Проверим достоверность отличия $\bar{P}^1$ и $\bar{P}$ с использованием критерия Смирнова. Для этого вычислим его пороговое значение

$$\bar{D}_\alpha = \sqrt{-\ln \frac{\alpha}{2} / T},$$

которое сравним с разностью $\bar{D} = |\bar{P}_0 - \bar{P}|$.

Исходная гипотеза H_0 подтверждается, если $\bar{D} > \bar{D}_\alpha$ и $\bar{P}_0 > \bar{P}$, в противном случае, т. е. при $\bar{P}_0 < \bar{P}$, она отвергается.

Анализ результатов вычислительных экспериментов. Сравним эффективность приведенных выше методик проверки гипотезы о распределениях случайных величин по данным вычислительных экспериментов.

Последовательности случайных наблюдений $V_1 = (x^i, i = \overline{1, n_1})$ и $V_2 = (x^i, i = \overline{1, n_2})$ 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 $[0, 1]$. При их формировании использовался стандартный датчик псевдослучайных величин среды визуального программирования Delphi.

Вычислительные эксперименты при фиксированных условиях исследования проводились 100 раз. По полученным результатам оценивалась вероятность $\bar{F}_0$ выполнения гипотезы H_0 о тождественности законов распределения случайных величин на основе исследуемых методик. Риск α отвергнуть гипотезу H_0 принимался равным 0,05.

Синтез непараметрического классификатора осуществлялся на основе параболических ядерных функции В. А. Епанечникова [9]. При формировании методики 2 значение $T = 10$. В критерии Пирсона использовалась формула Старджесса для разбиения области изменения исследуемых случайных величин на N интервалов.

Были получены следующие зависимости оценок вероятностей $\bar{F}_0$ выполнения гипотезы H_0 от объема экспериментальных данных при априори тождественных (табл. 1, 2) и разных (табл. 3) законах распределения случайных величин. В таблицах использованы следующие обозначения: КС – критерий Смирнова; М1, М2 – методики 1 и 2 соответственно; КП – критерий Пирсона.

Таблица 1

Зависимости оценок вероятностей $\bar{F}_0$ справедливости гипотезы H_0 от объема экспериментальных данных $n = n_1 + n_2$ при $n_1 = 1, 2n_2$ в условиях сравнения двух априори тождественных законов распределения случайных величин

n	Равномерные законы распределения				Нормальные законы распределения			
	КС	М1	М2	КП	КС	М1	М2	КП
22	0,97	0,98	0,99	0,98	1,00	0,98	0,98	0,98
66	0,96	0,95	0,96	0,94	0,98	0,96	0,98	0,98
110	0,98	0,95	0,96	0,99	0,99	0,99	1,00	0,97
154	0,96	0,92	0,97	0,96	0,96	0,93	0,97	0,97
198	0,98	0,93	0,96	0,92	0,99	0,97	1,00	0,96
242	0,97	0,96	0,97	0,95	0,96	0,96	0,96	0,91
286	0,96	0,92	0,94	0,98	0,96	0,94	0,95	0,96
330	0,96	0,94	0,95	0,97	0,97	0,94	0,98	0,97
374	1,00	0,93	0,97	0,99	0,96	0,93	0,95	0,95
418	0,97	0,93	0,95	0,98	0,97	0,93	0,96	0,96
462	0,96	0,91	0,94	0,94	0,96	0,9	0,94	0,95

Таблица 2

Зависимости оценок вероятностей $\bar{F}_0$ справедливости гипотезы H_0 от объема экспериментальных данных $n = n_1 + n_2$ при $n_1 = 2n_2$ в условиях сравнения двух априори тождественных законов распределения случайных величин

n	Равномерные законы распределения				Нормальные законы распределения			
	КС	M1	M2	КП	КС	M1	M2	КП
30	0,99	0,98	1,00	0,98	0,99	0,98	1,00	0,94
90	0,98	0,93	0,99	0,99	0,98	0,92	0,99	0,95
150	0,96	0,86	0,97	0,95	0,97	0,87	0,99	0,98
210	0,98	0,82	0,96	0,96	0,97	0,73	0,99	0,94
270	1,00	0,77	0,99	0,99	0,96	0,69	1,00	0,97
330	0,93	0,70	0,97	0,91	0,98	0,65	0,99	0,97
390	0,94	0,74	0,96	0,96	0,97	0,63	0,97	0,94
450	0,96	0,67	0,97	0,97	0,94	0,61	0,98	0,95

Таблица 3

Зависимости оценок вероятностей $\bar{F}_0$ справедливости гипотезы H_0 от n в условиях сравнения нормального и равномерного законов распределения случайных величин

$n = n_1 + n_2$ при $n_1 = 1,2n_2$					$n = n_1 + n_2$ при $n_1 = 2n_2$				
n	КС	M1	M2	КП	n	КС	M1	M2	КП
22	0,84	0,76	0,93	0,65	30	0,9	0,45	0,89	0,53
66	0,58	0,21	0,27	0,22	90	0,53	0	0,10	0,02
110	0,23	0,01	0,03	0,01	150	0,17	0	0,02	0
154	0,14	0	0	0,01	210	0,03	0	0	0
198	0	0	0	0	270	0	0	0	0

Анализ данных табл. 1 показывает, что если сравниваемые законы распределения случайных величин априори тождественны, то эффективность критериев Смирнова и Пирсона и методики 2 достоверно не отличается. Показатели методики 1 сопоставимы с ними при малых объемах обучающей выборки ($n < 100$). Однако эффективность этой методики существенно снижается с ростом отношения $\frac{n_1}{n_2}$ ($n_1 > n_2$) и значений n .

При анализе последовательностей случайных величин с равномерным и нормальным законами распределения в условиях $n > 200$ сравниваемые критерии безошибочно отклоняют гипотезу H_0 (см. табл. 2). При $n < 30$ результаты их использования неудовлетворительны. В интервале $30 < n < 180$ применение критерия Смирнова нецелесообразно, так как ему свойственна большая вероятность подтверждения гипотезы H_0 в условиях априори разных законов распределения случайных величин. При этом эффективность критерия Пирсона, методик 1 и 2 сопоставима и повышается с ростом отношения $\frac{n_1}{n_2}$.

Таким образом, применение принципов коллективного оценивания позволяет повысить эффективность методики 1 проверки гипотезы о распределениях случайных величин, основанной на непараметрическом алгоритме распознавания образов. Показатели методики 2 и критерия Пирсона сопоставимы.

Их применение по сравнению с критерием Смирнова является более предпочтительным в условиях малого объема анализируемых случайных последовательностей. Преимущество методики 2 состоит в обходе проблемы декомпозиции области значений случайных величин на интервалы, свойственной критерию Пирсона. Перспективность этой методики заключается в возможности ее обобщения на решение задач проверки гипотез о распределения многомерных случайных величин.

Библиографические ссылки

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. Непараметрические алгоритмы распознавания образов в задаче проверки статистической гипотезы о тождественности двух законов распределения случайных величин // Автоматизация. 2010. № 6. С. 47–53.
4. Лапко А. В., Лапко В. А. Применение непараметрического алгоритма распознавания образов в задаче проверки гипотезы о распределениях случайных величин // Системы упр. и информ. технологии. 2010. № 3 (41). С. 8–11.
5. Шаракшанэ А. С., Железнов И. Г., Ивницкий В. А. Сложные системы. М.: Высш. шк., 1977.
6. Непараметрические системы классификации / А. В. Лапко, В. А. Лапко, М. И. Соколов, С. В. Чен-

цов. Новосибирск : Сиб. изд. фирма СО РАН «Наука», 2000.

7. Parzen E. On Estimation of a Probability Density Function and Mode // Ann. Math. Statistic. 1962. Vol. 33, № 3. P. 1065–1076.

8. Лапко А. В., Лапко В. А. Анализ асимптотических свойств непараметрической оценки уравнения

разделяющей поверхности в дуальтернативной задаче распознавания образов // Автометрия. 2010. Т. 46, № 3. С. 48–53.

9. Епанечников В. А. Непараметрическая оценка многомерной плотности вероятности // Теория вероятности и ее применения. 1969. Т. 14. Вып. 1. С. 156–161.

A. V. Lapko, V. A. Lapko

COMPARISON OF NONPARAMETRIC TESTING CRITERIA OF HYPOTHESIS OF DISTRIBUTION OF RANDOM VARIABLES

From a perspective of collective estimation principles, the authors worked out the technique of testing, for a hypothesis of identity of laws of distributions of random variables, based on nonparametric algorithms of pattern recognition. Results of its comparison with Smirnov and Pearson criterion are presented.

Keywords: nonparametric statistics, pattern recognition, statistical hypothesis testing, random variables distribution, Smirnov criterion, Pearson criterion.

© Лапко А. В., Лапко В. А., 2011

УДК 62.501

А. В. Медведев

ТЕОРИЯ НЕПАРАМЕТРИЧЕСКИХ СИСТЕМ. АКТИВНЫЕ ПРОЦЕССЫ – I

Исследуется проблема моделирования организационных процессов и управления ими. Существенную роль при этом играет априорная информация. Обсуждаются вопросы принципиального отличия управления техническими и активными системами. Основной особенностью здесь является необходимость введения контуров управления в систему контроля некоторых выходных переменных, а также систему управления процессом реализации найденного управляющего воздействия.

Ключевые слова: активные системы, идентификация, априорная информация, измерение, гипотезы, непараметрические модели, дискретно-непрерывные процессы, К-модели.

Теория – в виду практики.

Девиз конгрессов IFAC

Ограничиваясь одними рассуждениями, мы уподобились бы некоторым древним философам, пытавшимся добывать законы природы из собственной головы. При этом неизбежно возникает опасность, что построенный таким образом мир при всех своих достоинствах окажется весьма мало похожим на действительный...

Л. Д. Ландау

В последние десятилетия внимание исследователей все больше привлекают процессы организационного характера. К ним относятся процессы, протекающие с участием человека или коллективов людей, в частности объектов промышленности, коммерческих структур, региональных образований и др. Характерными особенностями последних являются неполнота априорных данных, неопределенность, взаимосвязанность, трудность формирования согласованных целей и способов их достижения и др. Неполнота априорных сведений приводит к необходимости формулировать те или иные задачи локального характера

в различных, принципиально отличающихся постановках, а их объединение в единую систему представляет серьезные теоретические трудности. В частности, задача управления организацией, коллективами во многом остается в большей степени искусством, чем наукой. Обусловлено это тем, что присутствие человека (коллектива) в исследуемом процессе требует учета ряда факторов, а именно: морального, психологического, престижного, и других черт, свойственных человеку, наиболее важной из которых является искажение информации о своих возможностях, целях, способах и средствах их достижения.