ПОМОЩЬ ВЕДУЩИХ СПЕЦИАЛИСТОВ

УДК 681.513

Александр Васильевич ЛАПКО,
главный научный сотрудник Института вычислительного моделирования СО РАН (г.Красноярск), доктор технических наук, профессор, заслуженный деятель науки Российской Федерации

Василий Александрович ЛАПКО,
ведущий научный сотрудник Института вычислительного моделирования СО РАН (г.Красноярск), доктор технических наук, профессор

Вячеслав Витальевич МОЛОКОВ,
доцент кафедры социально-экономических наук и информатики Сибирского юридического института МВД России (г.Красноярск), кандидат технических наук, доцент

МЕТОДИКА ПРОВЕРКИ ГИПОТЕЗЫ О РАСПРЕДЕЛЕНИЯХ МНОГОМЕРНЫХ СЛУЧАЙНЫХ ВЕЛИЧИН, ОСНОВАННАЯ НА ИСПОЛЬЗОВАНИИ НЕПАРАМЕТРИЧЕСКИХ АЛГОРИТМОВ РАСПОЗНАВАНИЯ ОБРАЗОВ

TECHNIQUE OF TESTING HYPOTHESIS ABOUT THE MULTIDIMENSIONAL RANDOM VARIABLES DISTRIBUTIONS BASED ON THE USE OF NONPARAMETRIC IMAGES RECOGNITION ALGORITHMS

Предлагается методика проверки гипотезы о тождественности законов распределения многомерных случайных величин, основанная на использовании непараметрических алгоритмов распознавания образов и принципов коллективного оценивания.

The technique for testing of the hypothesis about identity of laws of multidimensional random variables distribution based on use of nonparametric images recognition algorithms and principles of collective estimation is offered.

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

Keywords: *nonparametric statistics, images recognition, hypothesis testing, distributions of random variables.*

Задачи проверки гипотезы о распределении случайных величин являются классическими при проведении всесторонних статистических исследований и широко распространены в научной практике. Не является исключением и область правоохранительной деятельности. Подобные подходы применимы для анализа закономерностей развития преступности, моделирования и прогнозирования ее показателей, выявления взаимосвязи в изучаемых процессах, классификации объектов или признаков, исследования криминологических характеристик преступного поведения и т.п.[1]

Относительно новым в общей теории статистики является применение непараметрических алгоритмов распознавания образов и принципов коллективного оценивания для проверки статистических гипотез. Одно из реализуемых направлений применения методов непараметрической статистики связано с оценкой эффективности деятельности экспертно-криминалистических подразделений органов внутренних дел. Исходные данные деятельности подразделений представлены статистическими показателями и содержат выборки наблюдений многомерной случайной величины. Проверка гипотезы о равенстве законов распределения исходных выборок может являться основой объединения их в группы классов, близких не столько по уровню и размерности величин, их характеризующих, сколько по степени связей между признаками и однородности законов формирования. Такая постановка задачи позволит выявить критерии оценки эффективности работы экспертно-криминалистических подразделений и согласовать существующие методики с результатами вычислительных экспериментов. Полученные данные могут являться основой для принятия различных управленческих решений руководством правоохранительных органов.

Для проверки гипотез о распределениях случайных величин широко используется критерий согласия К.Пирсона, который не зависит от распределений случайных величин и их размерности.[2] Однако методика формирования критерия Пирсона содержит трудно формализуемый этап разбиения области возможных значений случайной величины на многомерные интервалы. Данный этап отсутствует в критерии Колмогорова – Смирнова, который позволяет проверять гипотезы о распределениях одномерных случайных величин.[3]

В статье А.В.Лапко[4] 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 количество элементов сравниваемых последовательностей случайных величин отличаются незначительно. При различных объемах случайных последовательностей наблюдается снижение эффективности предлагаемой методики. Данный факт согласуется с результатами исследований[5], где показано значительное ухудшение аппроксимационных свойств непараметрической оценки уравнения разделяющей поверхности при увеличении степени неравномерности распределения элементов обучающей выборки между классами.

Цель данной работы состоит в развитии предлагаемой методики для решения задач проверки гипотез о распределениях многомерных случайных величин.

Работа выполнена при поддержке гранта ФЦП «Научные и научно-педагогические кадры инновационной России» на 2009–2013 гг., ГК №02.740.11.0621.

Модифицированная методика проверки гипотезы о распределениях случайных величин. Пусть X_1 и X_2 – две генеральные совокупности с произвольными законами распределения.

Необходимо по независимым выборкам $V_1 = \left(x^i, i = \overline{1, n_1}\right)$ и $V_2 = \left(x^i, i = \overline{1, n_2}\right)$ многомерных случайных величин $x = \left(x_v, v = \overline{1, k}\right)$, извлеченным 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 двуальтернативной 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 многомерных случайных величин к проверке гипотезы о равенстве статистической оценки вероятности ошибки распознавания образов значению 0.5.

Предлагаемая методика предполагает выполнение следующих действий.

1. Пусть количество элементов сравниваемых последовательностей случайных величин отличается значительно, например $n_1 > n_2$. Сформировать набор сравниваемых последовательностей $\left(V_1(j) = \left(x^i, i \in I_j\right), V_2 = \left(x^i, i = \overline{1, n_2}\right)\right)$, $j = \overline{1, T}$. 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 сравниваемую последовательность $V_1(j)$. Присвоим элементам множества I_j значения $n_2 + t$, $t = \overline{1, n_2}$.

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

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

– указание 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[6],

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

При формировании оценки уравнения разделяющей поверхности

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

будем использовать непараметрические оценки

$$\overline{p}_2(x) = \frac{1}{n_2} \sum_{i=1}^{n_2} \prod_{v=1}^{k} \frac{1}{c_v} \Phi\left(\frac{x_v - x_v^i}{c_v}\right),$$

$$\overline{p}_1^j(x) = \frac{1}{n_2} \sum_{i=n_2+1}^{2n_2} \prod_{v=1}^{k} \frac{1}{c_v} \Phi\left(\frac{x_v - x_v^i}{c_v}\right)$$

плотностей вероятности распределения многомерной случайной величины x в классах Ω_1, Ω_2 типа Розенблатта – Парзена.[7] Ядерные функции $\Phi(u_v)$ удовлетворяют условиям $\Phi(u_v) = \Phi(-u_v)$,

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

значения их коэффициентов размытости c_v убывают с ростом n_2.

Тогда статистика (2) представляется выражением

$$\tilde{f}_{12}^{j}(x)=\frac{1}{n_2}\sum_{i=1}^{2n_2}\sigma(i)\prod_{v=1}^{k}\frac{1}{c_v}\Phi\left(\frac{x_v-x_v^i}{c_v}\right). \quad (3)$$

Выбор оптимальных значений $\bar{c}_v, v=\overline{1,k}$ коэффициентов размытости $c=\left(c_v, v=\overline{1,k}\right)$ непараметрического решающего правила $\bar{m}_j(x)$ осуществляется из условия минимума оценки вероятности ошибки распознавания образов

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

где индикаторная функция

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

$\bar{\sigma}(t)$ – «решение» алгоритма $\bar{m}_j(x)$ о принадлежности значений x^t к тому либо иному классу Ω_1, Ω_2, полученное в соответствии с правилом (1).

При вычислении $\bar{\rho}_j(c)$ «решение» $\bar{\sigma}(t)$ алгоритма (1) определяется в соответствии со знаком статистики

$$\tilde{f}_{12}^{j}\left(x^t\right)=\frac{1}{n_2}\sum_{\substack{i=1\\ i\neq t}}^{2n_2}\sigma(i)\prod_{v=1}^{k}\frac{1}{c_v}\Phi\left(\frac{x_v^t-x_v^i}{c_v}\right),$$

то есть ситуация x^t, которая подается на контроль, исключается из процесса обучения.

4.Проверить гипотезу $\bar{H}_\circ(j)$: $\bar{\rho}_j(\bar{c})=0.5$ в соответствии с критерием Колмогорова. Для этого сравним его пороговое значение[8]

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

с отклонением $\bar{D}_{12}^{j}=\left|0.5-\bar{\rho}_j(\bar{c})\right|$. Здесь α – вероятность (риск) отвергнуть правильную гипотезу $\bar{H}_\circ(j)$.

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

5. В соответствии с пунктами 2–4 проверить гипотезы $\bar{H}_\circ(j)$ на основе последовательностей случайных величин $(V_1(j),V_2)$, $j=\overline{1,T}$. 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, а $\bar{S}$ – отклонения гипотез $\bar{H}_\circ(j)$, $j=\overline{1,T}$.

6. Проверить достоверность отличия $\bar{P}^1$ и $\bar{P}$ с использованием критерия Смирнова.

Для этого вычислим его пороговое значение

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

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

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.

Предлагаемая методика позволяет расширить условия применения критерия Колмогорова – Смирнова на задачи проверки гипотез о распределениях многомерных случайных величин. Ее использование обеспечивает обход проблемы разбиения области возможных значений случайной величины на многомерные интервалы, что свойственно критерию Пирсона.

Представленные результаты работы могут быть полезны при решении различных прикладных задач обработки экспериментальных данных в технических, социально-экономических, медико-биологических и иных системах.

1 Молоков В.В. Направления применения методов непараметрической статистики в решении задач профилактики и борьбы с преступностью // Актуальные проблемы борьбы с преступностью в Сибирском регионе : сборник материалов XII международной научно-практической конференции. Красноярск : СибЮИ МВД России, 2009. Ч.2. С.115-118.

2 Пугачев В.С. Теория вероятностей и математическая статистика. М.: Наука: Главная редакция физико-математической литературы, 1979.

3 Смирнов Н.В. Оценка расхождения между эмпирическими кривыми распределений в двух независимых выборках // Бюллетень МГУ. Сер. А. Вып.2. 1939. С.3–14.

4 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.

5 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.

6 Лапко А.В., Лапко В.А., Соколов М.И., Ченцов С.В. Непараметрические системы классификации. Новосибирск : 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 Шаракшанэ А.С., Железнов И.Г., Ивницкий В.А. Сложные системы. М.: Высш. шк., 1977.