УДК: 519.7

© 2020 г. А.В. Лапко^{1,2}, д-р техн. наук,

В.А. Лапко^{1,2}, д-р техн. наук

(¹Институт вычислительного моделирования СО РАН, Красноярск,

²Сибирский государственный университет науки и технологий
имени академика М.Ф. Решетнева, Красноярск)

НЕПАРАМЕТРИЧЕСКИЙ АЛГОРИТМ РАСПОЗНАВАНИЯ ОБРАЗОВ В ЗАДАЧЕ ПРОВЕРКИ ГИПОТЕЗЫ О ТОЖДЕСТВЕННОСТИ ЗАКОНОВ РАСПРЕДЕЛЕНИЯ МНОГОМЕРНЫХ СЛУЧАЙНЫХ ВЕЛИЧИН В УСЛОВИЯХ СТАТИСТИЧЕСКИХ ДАННЫХ БОЛЬШОГО ОБЪЕМА

Рассматривается новая методика проверки гипотез о распределениях многомерных статистических данных большого объема. Обосновывается возможность замены проверки гипотезы о тождественности двух законов распределений многомерных случайных величин на проверку гипотезы о равенстве ошибки распознавания образов значению 0,5. Для проверки этой гипотезы используется методика доверительного оценивания вероятности ошибки распознавания образов либо критерий Колмогорова. Обучающая выборка формируется по статистическим данным сравниваемых законов распределения. В условиях больших объемов статистических данных синтез непараметрического алгоритма распознавания образов осуществляется на основе регрессионных оценок плотностей вероятности распределения случайных величин в классах. Предлагаемые алгоритмы распознавания образов позволяют уменьшить объем обучающей выборки за счет декомпозиции области значений случайных величин. Рассматривается методика выбора оптимальных параметров декомпозиции области значений независимых случайных величин. Наиболее близким аналогом предлагаемого подхода является критерий Пирсона. Эффективность полученных результатов подтверждается их применением при анализе данных дистанционного зондирования.

Ключевые слова: проверка статистических гипотез, многомерные случайные величины, непараметрический алгоритм распознавания образов, ядерная оценка плотности вероятности, дискретизация области значений случайных величин, критерий Пирсона.

Введение

Задачи проверки гипотез о распределениях случайных величин являются актуальными при оценивании состояний объектов различной природы и выборе их информативных признаков. При проверке гипотез о распределении случайных величин используется критерий согласия Пирсона, который требует использования трудно формализуемого этапа разбиения области значений случайных величин на многомерные интервалы.

В статьях [1 – 3] обосновывается возможность использования непараметрических алгоритмов распознавания образов, соответствующих критерию максимального правдоподобия, в задачах проверки статистических гипотез о распределениях случайных величин. Результаты использования предлагаемой методики сопоставимы с критерием Колмогорова – Смирнова для одномерных случайных величин 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 [4 – 7], 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 предложенной методики на решение задач проверки гипотез о распределениях многомерных случайных величин в условиях больших объемов статистических данных, которые характерны при анализе, например, данных дистанционного зондирования.

Постановка задачи

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 $V_2 = (x^i, i = \overline{1, n_2})$ многомерных данных $x = (x_v, v = \overline{1, k})$, извлеченным из данных генеральных совокупностей, проверить гипотезу

$$H_0 : F_1(x) \equiv F_2(x),$$

о тождественности их функций распределения $F_1(x)$, $F_2(x)$ в условиях больших объемов n_1 , n_2 статистических данных.

Если при решении двухальтернативной задачи распознавания образов в соответствии с критерием максимального правдоподобия вероятность

ошибки классификации равна значению 0.5, то законы распределения случайных величин в области определения классов очевидно совпадают. Поэтому появляется возможность перехода от задачи сравнения законов распределения многомерных случайных величин к проверке гипотезы $\bar{H}_0$ о равенстве вероятности ошибки распознавания образов значению 0,5.

Непараметрические алгоритмы распознавания образов в условиях обучающих выборок большого объема

Сформируем на основании V_1 и V_2 обучающую выборку $V = (x^i, \sigma(i), i = \overline{1, n})$, где $n = n_1 + n_2$, а указания $\sigma(i)$ о принадлежности ситуации x^i к V_1 либо V_2 определяется соотношением

$$\sigma(i) = \begin{cases} -1, & \text{если } x^i \in V_1 \\ 1, & \text{если } x^i \in V_2. \end{cases} \quad (1)$$

В этом случае элементы массивов данных V_1, V_2 в терминах задачи распознавания образов образуют обучающую выборку их принадлежности к классам Ω_1, Ω_2 соответственно.

По обучающей выборке V осуществим синтез непараметрического алгоритма распознавания образов, соответствующего критерию максимального правдоподобия [8, 9]:

$$\bar{m}(x) = \begin{cases} x \in \Omega_1, & \text{если } \bar{f}_{12}(x) \leq 0 \\ x \in \Omega_2, & \text{если } \bar{f}_{12}(x) > 0, \end{cases} \quad (2)$$

где оценка уравнения разделяющей поверхности между классами определяется выражением

$$\bar{f}_{12}(x) = \bar{p}_2(x) - \bar{p}_1(x). \quad (3)$$

В условиях больших объемов обучающей выборки V возрастают временные затраты при решении задачи распознавания образов с использованием традиционных непараметрических оценок плотностей вероятностей $\bar{p}_1(x), \bar{p}_2(x)$ ядерного типа.

Чтобы обойти возникающую проблему при оценивании плотностей вероятностей случайных величин в классах, будем использовать их регрессионные оценки [10, 11]. Синтез подобной оценки плотности вероятности основан на «сжатии» исходных выборок V_1, V_2 в массивы данных

$$\bar{V}_t = (z_1^j, \dots, z_k^j, \bar{P}^j, j = \overline{1, N_t}), \quad t = 1, 2$$

путем декомпозиции области значений случайных величин $x = (x_1, \dots, x_k)$ на

многомерные интервалы, где N_t – количество многомерных интервалов в области классов; $z^j = (z_1^j, \dots, z_k^j)$ – центры интервалов дискретизации; $\bar{P}^j$ – частоты принадлежности случайных величин из массивов данных V_t , $t=1, 2$ этим интервалам. Объем N_t полученных массивов данных значительно меньше объема n_t исходных данных V_t , $t=1, 2$.

В качестве оценки плотности вероятности $p_t(x)$ примем статистику [11]

$$\bar{p}_t(x_1, \dots, x_k) = \frac{1}{\prod_{v=1}^k c_v} \sum_{j=1}^{N_t} \bar{P}^j \prod_{v=1}^k \Phi\left(\frac{x_v - z_v^j}{c_v}\right), \quad t = \overline{1, 2}, \quad (4)$$

где ядерные функции $\Phi(u_v)$ удовлетворяют условиям:

$$\begin{aligned} \Phi(u_v) &= \Phi(-u_v), \quad 0 \leq \Phi(u_v) < \infty, \quad \int_{-\infty}^{\infty} \Phi(u_v) du_v = 1, \quad \int_{-\infty}^{\infty} u_v^2 \Phi(u_v) du_v = 1, \\ \int_{-\infty}^{\infty} u_v^m \Phi(u_v) du_v &< \infty, \quad 0 \leq m < \infty, \quad v = \overline{1, k}. \end{aligned}$$

Коэффициенты размытости c_v ядерных функций убывают с ростом N_t , $t=1, 2$.

Значения коэффициентов размытости c_v , $v = \overline{1, k}$ ядерных функций зависят от интервалов значений случайных величин x_v , $v = \overline{1, k}$. С ростом длины интервалов изменения случайных величин значения коэффициентов размытости увеличиваются. Поэтому будем полагать, что $c_v = c \sigma_v$, $v = \overline{1, k}$, где σ_v – среднее квадратическое отклонение компонент случайных величин x_v , $v = \overline{1, k}$.

Тогда непараметрическая оценка $\bar{f}_{12}(x_1, \dots, x_k)$, восстанавливаемая по данным $\bar{V}_t$, $t=1, 2$, запишется в виде

$$\bar{f}_{12}(x_1, \dots, x_k) = \left(c^k \prod_{v=1}^k \sigma_v \right)^{-1} \sum_{j=1}^2 (-1)^j \sum_{i=1}^{N_t} \bar{P}_t^i \prod_{v=1}^k \Phi\left(\frac{x_v - z_v^i}{c \sigma_v}\right). \quad (5)$$

Соответствующее этому уравнению разделяющей поверхности решающее правило определяется соотношением

$$\bar{m}(x) = \begin{cases} x \in \Omega_1, & \text{если } \bar{f}_{12}(x) \leq 0 \\ x \in \Omega_2, & \text{если } \bar{f}_{12}(x) > 0. \end{cases} \quad (6)$$

Предлагаемая регрессионная оценка многомерного уравнения разделя-

ющей поверхности между классами Ω_1, Ω_2 обладает свойствами асимптотической несмещенности и состоятельности [11].

Оптимизация непараметрической оценки уравнения разделяющей поверхности между классами

При решении данной задачи возникают две проблемы, которые связаны с выбором количества интервалов дискретизации области значений многомерной случайной величины в классах и определением оптимальных коэффициентов размытости в непараметрической оценке уравнения разделяющей поверхности (5) из условия минимума оценки вероятности ошибки распознавания образов.

В работах [12, 13] предложена методика расчета оптимального количества интервалов дискретизации области значений многомерной случайной величины с плотностью вероятности $p(x_1, \dots, x_k)$

$$N_k^* = \left(\alpha(k) n \prod_{v=1}^k \Delta_v \int_{-\infty}^{+\infty} \dots \int_{-\infty}^{+\infty} p^2(x_1, \dots, x_k) dx_1 \dots dx_k \right)^{1/2} = \alpha_k \sqrt{n}, \quad (7)$$

где Δ_v – длина интервала значений x_v , $v = \overline{1, k}$. Коэффициент $\alpha(k) = (2k-1)/k^2 \leq 1$ и его значения уменьшаются с ростом размерности k случайной величины. Установлено, что коэффициент $\alpha(k)$ определяется только видом плотности вероятности и не зависит от ее параметров.

Предположим, что случайные величины x_v , $v = \overline{1, k}$ являются независимыми. Тогда выражение (7) запишется в виде

$$N_k^* = \left(\prod_{v=1}^k \left(\left(\frac{2k-1}{k^2} n \right)^{\frac{1}{k}} \Delta_v \int_{-\infty}^{+\infty} p^2(x_v) dx_v \right) \right)^{1/2} = \prod_{v=1}^k N^*(v),$$

где

$$N^*(v) = \left(\left(\frac{2k-1}{k^2} n \right)^{\frac{1}{k}} \Delta_v \int_{-\infty}^{+\infty} p^2(x_v) dx_v \right)^{1/2}. \quad (8)$$

В работе [14] показано, что для одномерного случая выражение $W_2 = \left(\Delta_v \int_{-\infty}^{+\infty} p^2(x_v) dx_v \right)^{1/2}$ является константой, которая определяется видом плотности вероятности и не зависит от ее параметров. Для семейства одномерных плотностей вероятностей эта константа допускает оценивание по

значению коэффициента контрэксцесса

$$\overline{W}_2(\delta_v) = 31,27\delta_v^4 - 62,73\delta_v^3 + 46,34\delta_v^2 - 16,18\delta_v + 3,8 \quad (9)$$

либо

$$\overline{W}_2(\delta_v) = 1,06\delta_v^{-0,386}. \quad (10)$$

В этих условиях средняя относительная ошибка аппроксимации для модели (9) определяется значением 0,0275, а для модели (10) – 0,037. Оценивание коэффициента контрэксцесса δ_v осуществляется по каждой компоненте x_v , $v = \overline{1, k}$.

Тогда методика дискретизации области значений случайной величины $x = (x_v, v = \overline{1, k})$ в каждом классе на многомерные интервалы предполагает выполнение следующих действий:

1. По статистическим данным $V_t = (x^i, i = \overline{1, n})$, характеризующихся плотностью вероятности $p_t(x)$ распределения случайных величин $x = (x_1, \dots, x_k)$ в классе Ω_t , определить оценки $\overline{\Delta}_v$ длин интервалов изменения значений случайных величин как разности между минимальными и максимальными значениями x_v^i , $i = \overline{1, n}$ и оценить соответствующие им коэффициенты контрэксцесса

$$\overline{\delta}_v = 1/\sqrt{\overline{\eta}_v},$$

где

$$\overline{\eta}_v = \frac{\frac{1}{n} \sum_{i=1}^n (x_v^i - \overline{x}_v)^4}{\left(\frac{1}{n} \sum_{i=1}^n (x_v^i - \overline{x}_v)^2 \right)^2}, \quad \overline{x}_v = \frac{1}{n} \sum_{i=1}^n x_v^i, \quad v = \overline{1, k}.$$

2. Используя формулу (9) либо (10), вычислить оценки констант $\overline{W}_2(\overline{\delta}_v)$, $v = \overline{1, k}$ для каждого класса.

3. По значениям $\overline{\Delta}_v$, $\overline{W}_2(\overline{\delta}_v)$, $v = \overline{1, k}$, n и k в соответствии с выражением (8) определить количество интервалов дискретизации случайных величин в каждом классе

$$\overline{N}^*(v) = \left[\left(\frac{2k-1}{k^2} n \right)^{\frac{1}{k}} \overline{W}_2(\overline{\delta}_v) \right]^{1/2}$$

и их длину

$$\overline{\overline{\Delta}}_v = \overline{\Delta}_v / \overline{N}^*(v), \quad v = \overline{1, k}.$$

Используя данные рекомендации, осуществим декомпозицию области

значений случайных величин $x = (x_1, \dots, x_k)$ в классах Ω_1, Ω_2 по данным V_1, V_2 .

Для упрощения изложения методики декомпозиции области значений многомерной случайной величины принадлежность параметров $\bar{\delta}_v, \bar{\eta}_v, \bar{N}^*(v), \bar{\Delta}_v, v = \overline{1, k}$ законам распределения $p_t(x_1, \dots, x_k), v = \overline{1, k}$ не указывается. Их принадлежность классам Ω_1, Ω_2 следует из логики изложения методики.

Первый этап оптимизации непараметрического алгоритма распознавания образов определяет его обучающую выборку $\bar{V} = (z^j, \bar{P}^j, \bar{\sigma}(j), j = \overline{1, N})$, которая является основой при оптимизации решающего правила (6) по коэффициентам размытости ядерных функций.

Оптимизация непараметрического решающего правила (6) по параметру c коэффициента размытости ядерных функций осуществляется в режиме «скользящего экзамена» из условия минимума статистической оценки вероятности ошибки распознавания образов

$$\bar{\rho}(c) = \frac{1}{N} \sum_{t=1}^N 1(\sigma(t), \bar{\sigma}(t)), \quad (11)$$

$$1(\sigma(t), \bar{\sigma}(t)) = \begin{cases} 0, & \text{если } \sigma(t) = \bar{\sigma}(t) \\ 1, & \text{если } \sigma(t) \neq \bar{\sigma}(t), \end{cases}$$

где $N = N_1 + N_2$; $\sigma(t), \bar{\sigma}(t)$ – соответственно «указания учителя» и решение алгоритма (6) о принадлежности ситуации z^t к одному из двух классов. При формировании решения $\bar{\sigma}(t)$ ситуация z^t исключается из процесса обучения в непараметрической статистике (5).

В этом случае принадлежность ситуации $(x_1^t, \dots, x_k^t)$ к одному из двух классов осуществляется по знаку непараметрической оценки уравнения разделяющей поверхности

$$\bar{f}_{12}(x_1^t, \dots, x_k^t) = \left(c^k \prod_{v=1}^k \sigma_v \right)^{-1} \sum_{j=1}^2 (-1)^j \sum_{\substack{i=1 \\ i \neq t}}^{N_j} \bar{P}_j^i \prod_{v=1}^k \Phi \left(\frac{x_v^t - z_v^i}{c \sigma_v} \right).$$

Методика проверки гипотез о распределениях многомерных случайных величин

1. Следуя рекомендациям «сжатия» исходных статистических данных V_1, V_2 , сформировать обучающую выборку $\bar{V} = \bar{V}_1 \cup \bar{V}_2$ для решения задачи распознавания классов Ω_1, Ω_2 для проверки гипотез $\bar{H}_0$ об их тождественности.

2. Осуществить синтез непараметрического алгоритма распознавания

образов (6) с использованием регрессионных оценок плотностей вероятности распределения случайных величин в классах Ω_1 , Ω_2 и определить оценку вероятности ошибки распознавания образов $\bar{\rho}$ в соответствии с выражением (11).

3. Проверить гипотезу $\bar{H}_0 : \bar{\rho} = 0,5$, используя методику доверительного оценивания вероятности ошибки распознавания образов ρ [15] либо критерий Колмогорова [16]. В первом случае гипотеза $\bar{H}_0$ выполняется, если значение $\rho = 0,5$ попадает в интервал

$$[\rho_1, \rho_2] = \frac{\bar{\rho} + \frac{t_\alpha^2}{2n} \mp t_\alpha \sqrt{\frac{\bar{\rho}(1-\bar{\rho})}{n} + \frac{t_\alpha^2}{4n^2}}}{1 + \frac{t_\alpha^2}{n}}.$$

Здесь t_α значение аргумента функции Лапласа, при котором выполняется условие

$$2F(t_\alpha) = \alpha,$$

где α – коэффициент доверия принадлежности значения ρ интервалу $[\rho_1, \rho_2]$.

При использовании критерия Колмогорова отклонение $\bar{D}_{12} = |0,5 - \bar{\rho}(\bar{c})|$ сравнивается с его пороговым значением

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

где β – вероятность (риск) отвергнуть гипотезу $\bar{H}_0$.

Если выполняется соотношение $\bar{D}_{12} < D_\beta$, то гипотеза $\bar{H}_0$ справедлива и риск ее отвергнуть равен β .

Предложенная методика была использована при анализе данных дистанционного зондирования лесных массивов территории юга Красноярского края. Проверялась гипотеза о тождественности законов распределения спектральных данных соответствующих темнохвойному лесу. Каждый элемент земной поверхности характеризовался шестью спектральными признаками $x = (x_1, \dots, x_6)$, которым свойственны следующие длины волн (нанометры): 450–515 (x_1), 525–600 (x_2), 630–680 (x_3), 845–885 (x_4), 1560–1660 (x_5), 2100–2300 (x_6). Исходная статистическая информация $V = (x_v^i, v = \overline{1, 6}, i = \overline{1, n})$ при $n=968$ при организации вычислительных эксперимента разделялась на две части $V(1)$, $V(2)$. При этом значения $\alpha=0,95$, а $\beta=0,05$. Применение предлагаемой методики позволяет сократить объем анализируемых данных в 4 раза. С увеличением объема n исходных статистических данных это отличие будет

значительно увеличиваться. Оценка вероятности отвергнуть гипотезу H_0 не превышает значения 0,03 при равных объемах статистических данных $V(1)$, $V(2)$, что незначительно больше соответствующего значения 0,02, полученного в работе [3]. Анализ результатов вычислительных экспериментов показывает, что при значительном отличии объемов n_1 , n_2 сравниваемых последовательностей V_1 , V_2 эффективность предлагаемого метода снижается. Решение данной проблемы может быть осуществлено путем использования принципов коллективного оценивания [17 – 19].

Заключение

Проверка гипотезы о тождественности двух многомерных законов распределения может быть заменена на проверку гипотезы о равенстве ошибки распознавания образов значению 0,5. Для этого необходимо сформировать обучающую выборку для решения задачи классификации по исходным статистическим данным. Для сокращения временных затрат на проверку гипотезы о тождественности многомерных случайных величин в условиях статистических данных большого объема целесообразно использовать регрессионные оценки плотности вероятности распределения случайных величин в классах. Их синтез основан на декомпозиции области значений случайных величин, что позволяет сократить объем исходных статистических данных при незначительном снижении эффективности принимаемых решений.

Перспективность исследований в данном направлении состоит в сравнении эффективности предлагаемого подхода с традиционным критерием Пирсона на основе предлагаемой методики декомпозиции области значений случайных величин на многомерные интервалы.

ЛИТЕРАТУРА

1. Лапко А.В., Лапко В.А. Сравнение эмпирической и теоретической функций распределения случайной величины на основе непараметрического классификатора // Автометрия. – 2012. – Т. 48, № 1. – С. 45-49.
2. Лапко А.В., Лапко В.А. Непараметрические алгоритмы распознавания образов в задаче проверки статистической гипотезы о тождественности двух законов распределения случайных величин // Автометрия. – 2010. – Т. 46, № 6. – С. 47-53.
3. Лапко А.В., Лапко В.А. Методика проверки гипотез о распределениях многомерных спектральных данных с использованием непараметрического алгоритма распознавания образов // Компьютерная оптика. – 2019. – Т. 43, № 2. – С. 238-244.
4. Лапко А.В., Лапко В.А. Анализ асимптотических свойств непараметрической оценки уравнения разделяющей поверхности в двувальтернативной задаче распознавания образов // Автометрия. – 2010. – Т. 46, № 3. – С. 48-53.

5. *Лапко А.В., Лапко В.А.* Асимптотические свойства многомерной непараметрической оценки уравнения разделяющей поверхности в дуальтернативной задаче распознавания образов // Системы управления и информационные технологии. – 2010. – Т. 39, № 1. – С. 16-19.
6. *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. – Т. 42, № 4. – С. 58-61.
7. *Лапко А.В., Лапко В.А.* Свойства непараметрической оценки многомерной решающей функции в алгоритме распознавания образов, соответствующего критерию максимума апостериорной вероятности // Информатика и системы управления. – 2019. – Т. 60, № 2. – С. 99-107. DOI: 10.22250/isu.2019.60.99-107
8. *Fukunaga K.* Introduction to statistical pattern recognition. – San Diego: Academic Press, 1990.
9. *Лапко А.В., Лапко В.А., Соколов М.И., Ченцов С.В.* Непараметрические системы классификации. – Новосибирск: Наука, 2000.
10. *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.
11. *Лапко А.В., Лапко В.А.* Регрессионная оценка многомерной плотности вероятности и ее свойства // Автометрия. – 2014. – Т. 50, №2. – С. 50-56
12. *Лапко А.В., Лапко В.А.* Выбор оптимального количества интервалов дискретизации области значений двумерной случайной величины // Измерительная техника. – 2016. – № 2. – С. 14-17.
13. *Лапко А.В., Лапко В.А.* Метод дискретизации области значений многомерной случайной величины // Измерительная техника. – 2019. – № 1. – С. 16-20.
14. *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.
15. *Пугачев В.С.* Теория вероятностей и математическая статистика: учебное пособие. – М.: Физматлит, 2002.
16. *Шаракшанэ А.С., Железнов И.Г., Ивницкий В.А.* Сложные системы. – М.: Высш. шк., 1977.
17. *Лапко В.А., Капустин А.Н.* Синтез нелинейных непараметрических решающих правил в задачах распознавания образов // Автометрия. – 2006. – Т. 42, № 6. – С. 26-33.
18. *Лапко А.В., Лапко В.А.* Разработка и исследование двухуровневых непараметрических систем классификации // Автометрия. – 2010. – Т. 46, № 1. – С. 70-78.
19. *Лапко А.В., Лапко В.А.* Синтез структуры семейства непараметрических решающих функций в задаче распознавания образов // Автометрия. – 2011. – Т. 47, № 4. – С. 76-82.

E-mail:

Лапко Александр Васильевич – lapko@ict.krasn.ru;

Лапко Василий Александрович – lapko@ict.krasn.ru.