УДК 519.2

А. В. Овсянников, доц., канд. техн. наук (БГУ, г. Минск);
О. Г. Барашко, доц., канд. техн. наук (БГТУ, г. Минск)

ФИЛЬТРАЦИЯ ГИСТОГРАММНОЙ ОЦЕНКИ ПЛОТНОСТИ ВЕРОЯТНОСТИ НА ОСНОВЕИ НЕЧЕТКОЙ ПРИНАДЛЕЖНОСТИ ДАННЫХ ИНТЕРВАЛУ ГРУППИРОВАНИЯ

Гистограммная оценка плотности распределения вероятности (ПВ) – одна из самых распространенных, исторически первых и общепринятых элементов описательной, прикладной статистики, в частности, в сфере статистической радиотехники, задачах ЭМС. Проблематика. связанная с гистограммной оценкой (ГОСТ Р 50.1.033-2001 Прикладная статистика) в прикладной статистике. также хорошо известна. В стационарном случае. при исследовании стационарных временных рядов. эффективность гистограммной оценки связана лишь с наличием достаточного времени для ее получения. Если же исследуется нестационарный объект или нестационарный временной ряд и требуется более точное распознавание момента времени. когда состояние объекта или динамика ряда, характеризуемая распределением, значимо изменилась. то построение простой, быстрой и эффективной гистограммной оценки ПВ является актуальным.

Развитие непараметрических методов и общая направленность тематики работ в области непараметрических процедур оценивания ПВ (ядерная. проекционная. сплайн и др.) не исключает применения гистограммных оценок. например переменных. к числу которых относятся и полиграммные. Одной из проблем «правильного» распознавания ПВ. наряду с выбором количества интервалов группирования данных, является возможная «изрезанность» гистограммы, что особенно характерно при относительно небольших наборах данных. Решение этой проблемы заключается в применении гистограммных фильтров, например усредняющего, медианного, гауссовского и др. Однако, их применение интуитивно и исходит в основном из практической целесообразности. В представляемой работе предлагается теоретически обоснованная методика построения гистограммного фильтра, учитывающая следующие соображения.

Во-первых. можно отказаться от строгой единичной функции включения данных в конкретный интервал группирования. Данные могут находиться вблизи границ интервала и при изменении числа интервалов оказаться в другом интервале. Расположение данных на интервале группирования (близость или удаленность от границ интервала),

может интерпретироваться как нечеткая принадлежность данных конкретному интервалу.

Во-вторых. введение понятия нечеткой принадлежности данных интервалу группирования позволяет за счет весовых функций перегруппировать эти данные так, чтобы уменьшилась «изрезанность» гистограммы и тем самым обеспечивалась ее сглаженность.

Для построения гистограммного сглаживающего фильтра и установления качественных свойств взвешенной гистограммной оценки определим *коэффициент статистической взаимосвязи* между числом v_j и числом $u_{Aj} = \sum_{x_i \in A_j} \mu_j(x_i)$. которое соответствует взвешенному числу данных попавших в *j*-тый интервал. Площадь столбца *j*-того интервала при стандартной гистограммной оценке составляет величину $v_j \Delta_x$, а площадь такого же столбца в случае ВГО составит $u_{Aj}\Delta_x = \sum_{x_i \in A_j} \mu_j(x_i)\Delta_x$. Отношение этих площадей при асимптотически больших значениях v_j и u_{Aj} (большом покрытии данными интервала A_j) будет стремиться к устойчивому соотношению в виде коэффициента статистической взаимосвязи

$$k = \lim_{i\to\infty} \frac{\sum_i \mu_j(x_i)}{\sum_i I_j(x_i)} = \frac{\overline{u}_{Aj}}{\overline{v}_j} = \frac{1}{\Delta_x} \int_{\Delta_x} \mu_j(x)dx . \qquad (1)$$

Черта над символом в формуле (1) означает, что эти числа получены при асимптотически большом покрытии данными интервала A_j и их отношение стремится к пределу в виде коэффициента *k*. Тот же результат можно получить в общем случае, используя усреднение функции принадлежности

$$k = \overline{u}_{Aj} / \overline{v}_j = \int_{\Delta_x} \mu f dx / \int_{\Delta_x} f dx .$$

Ограничиваясь первым членом разложения в ряд Тейлора ПВ в точке середины интервала группирования x_j, получим результат, совпадающий с (1). Формула (1) справедлива для внутренних интервалов гистограммы ($j = \overline{2, m-1}$). Для оконечных интервалов $j = 1, m$, исходя из условий нормировки (2) и в том случае, если $\Delta_\mu = x_{j+1} - x_{j-1}$, получим

$$k_e = \Delta_x^{-1} \left(\int_{\Delta_x/2} 1(x)dx + \int_{\Delta_x/2} \mu(x)dx \right) = (1+k)/2 . \qquad (2)$$

С учетом определенных формулами (1), (2) коэффициентов. k, k_e можем записать соотношение между числами $\overline{v}_j$ и $\overline{u}_j$, которое будет представлять собой *гистограммный фильтр нулевого порядка* в одномерном случае

$$\begin{cases} \overline{u}_j = \alpha\overline{v}_{j-1} + k^0\overline{v}_j + \alpha\overline{v}_{j+1},\ j = \overline{2,m-1}\ , \\ \alpha = \overline{u}_{Aj-1} / \overline{v}_{j-1} = \overline{u}_{Aj+1} / \overline{v}_{j+1} = (1-k)/2, \\ \overline{u}_1 = k_e\overline{v}_1 + (1-k_e)\overline{v}_2,\ \ \overline{u}_m = (1-k_e)\overline{v}_{m-1} + k_e\overline{v}_m. \end{cases} \tag{3}$$

Заменой переменных $\overline{v}_j$, $\overline{u}_j$ на $g_j = v_j / \Delta_x n$, и f_j^*, получим гистограммный фильтр относительно значений ПВ на интервале группирования данных

$$\begin{cases} f_j^* = \alpha g_{j-1} + k g_j + \alpha g_{j+1},\ j = \overline{2,m-1} \\ f_1^* = k_e g_1 + (1-k_e) g_2,\ f_m^* = (1-k_e) g_{m-1} + k_e g_m. \end{cases} \tag{4}$$

Введением итеративной процедуры для формул (3). (4) можно добиться большей степени сглаживания

$$\begin{cases} \overline{u}_j^{q+1} = \alpha\overline{u}_{j-1}^q + k\overline{u}_j^q + \alpha\overline{u}_{j+1}^q,\ j = \overline{2,m-1} \\ \overline{u}_1^{q+1} = k_e\overline{u}_1^q + (1-k_e)\overline{u}_2^q,\ \ \overline{u}_m^{q+1} = (1-k_e)\overline{u}_{m-1}^q + k_e\overline{u}_m^q, \end{cases} \tag{5}$$

$$\begin{cases} f_j^{*q+1} = \alpha\ f_{j-1}^q + k\ f_j^{*q} + \alpha f_{j+1}^{*q},\ j = \overline{2,m-1} \\ f_1^{*q+1} = k_e f_1^{*q} + (1-k_e) f_2^{*q},\ f_m^{*q+1} = (1-k_e) f_{m-1}^{*q} + k_e f_m^{*q}, \end{cases} \tag{6}$$

где q – порядковый номер итерации, q=1 соответствует процедуре (3). (4).

Используя аналогичный подход в фильтрации многомерных данных, в частности, гистограмм изображений, получаем теоретически обоснованные результаты.