УДК 519.2

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

ГИСТОГРАММНАЯ ФИЛЬТРАЦИЯ ДАННЫХ

Основная идея гистограммной фильтрации состоит в отказе от привычной для гистограммного оценивания единичной функции включения. Эта функция заменяется нечеткой функцией принадлежности данных конкретному интервалу группирования. В работе показано, что в этом случае можно получить методику построения семейства гистограммных фильтров данных, работающих эффективно на малом объеме данных, устраняющих сильную изрезанность гистограммы, уменьшающих зависимость формы гистограммы от выбираемого числа интервалов группирования и приводящих к «правильную» идентификации закона распределения.

Первый тип гистограммного фильтра реализуется алгоритмом

$$\begin{cases} u_j = \sum\limits_{t=j-1}^{j+1} \mu_{j,t} v_t, \qquad j = \overline{2, m-1}\ , \\ u_j = \mu_{j,j} v_j + \mu_{j,t} v_t, \quad j = 1, m, \quad t = (j-m)(m-3)/(m-1) + (m-1), \end{cases} \tag{1}$$

где v_j – число данных попавших в j-тый интервал группирования, $\mu_{j,t}$ – дискретный вариант функции принадлежности, m – число интервалов группирования. Эти уравнения должны рассматриваться с учетом дополнительных ограничениях на ступенчатые весовые функции – коэффициенты гистограммного фильтра

$$\begin{cases} \sum_{t=j-1}^{j+1} \mu_{j,t} = 1, \quad j = \overline{2, m\text{-}1}, \\ \mu_{j,j} + \mu_{j,(j-m)(m-3)/(m-1)+(m-1)} = 1, \quad j = 1, m. \end{cases} \tag{2}$$

$$\begin{cases} \sum_{t=j-1}^{j+1} \mu_{t,j} = 1, \quad j = \overline{2, m\text{-}1}, \\ \mu_{j,j} + \mu_{(j-m)(m-3)/(m-1)+(m-1),j} = 1, \quad j = 1, m. \end{cases} \tag{3}$$

Алгоритм (1)-(3) может быть охарактеризован как алгоритм обобщенного сглаживания (фильтрации) гистограммы. В отличии от тривиального равномерного сглаживания с весом 1/3, алгоритм (1)-(3) использует настраиваемые веса, определяющиеся априорной информацией о предполагаемом законе распределения. Причем эти веса определяются не заранее, а в процессе обработки данных.

Коэффициенты $\mu_{j,t}$ образуют матрицу $\mathbf{M} = \{\mu_{j,t}\}$ весовых коэффициентов фильтра размерностью $(m \times m)$. Ненулевыми элементами матрицы **M** являются только элементы, находящиеся на главной диагонали и соседних с ней – выше и ниже. Уравнения (2), (3) представляют собой уравнения ограничений на сумму значений коэффициентов $\mu_{j,t}$ по строкам и по столбцам матрицы **M**.

Для расчета неизвестных коэффициентов матрицы **M,** уравнений (2), (3) не достаточно. Требуется дополнительная информация о связи коэффициентов с априорной информацией относительно предполагаемого закона распределения.

Алгоритмически такую связь можно представить в виде преобразования, которое отображает профиль участка любого закона распределения в некий постоянный уровень

$$\begin{cases} \rho_j = \mu_{j,t} p_t, \\ j = \overline{2, m-1}, \qquad t = \overline{j-1, j+1} \\ j = 1, m, \qquad t = j, (j-m)(m-3)/(m-1) + (m-1). \end{cases} \tag{4}$$

Второй тип гистограммного фильтра реализует идею преобразования (4), но распространённого на все интервалы группирования данных

$$\begin{cases} \rho = \mu_j p_j, \qquad j = \overline{1, m}, \\ \sum_{j=1}^{m} \mu_j = 1 \end{cases}$$

Целесообразность применение полученных теоретических результатов для быстрой (на малых объемах данных) и эффективной идентификации изменяющихся законов распределения в описательной статистике, при обработке гистограмм изображений. Предложенная методика и разработанный алгоритм гистограммного фильтра легко встраивается в существующие алгоритмы построения гистограмм, например, функции hist, histfit платформы Matlab.

ЛИТЕРАТУРА

1. Орлов. Ю.Н. Оптимальное разбиение гистограммы для оценивания выборочной плотности функции распределения нестационарного временного ряда. *Препринты ИПМ им. М.В.Келдыша*. 2013. № 14. 26с. URL: http://library.keldysh.ru/preprint.asp?id =2013–14

2. Chong Gu, Yongho Jeon and Yi Lin. *Nonparametric density estimation in high-dimensions*. Statistica Sinica 23 (2013), 1131–1153.