УДК 519.2

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

ГИСТОГРАММНЫЙ ФИЛЬТР С НАСТРОЙКОЙ ПАРАМЕТРА СГЛАЖИВАНИЯ

Цель работы построить гистограммный фильтр, эффективно работающий на небольшом количестве данных, устраняющий изрезанность гистограммы в этом случае, ослабляющий зависимость формы гистограммы от числа интервалов группирования данных и дающий «правильную» идентификацию закона распределения.

Проблематика гистограммных оценок плотности хорошо известна: отсутствие единых взглядов на определение числа интервалов группирования данных (ГОСТ Р 50.1.033-2001 Прикладная статистика) и сильная изрезанность гистограммы при относительно малом числе наблюдений [1, 2].

Устранение этих проблем заключается в применении гистограммных фильтров, например, усредняющего, медианного, гауссовского и др. В то же время, их применение эмпирически интуитивно и исходит в основном из практической целесообразности. В работе предлагается теоретически обоснованная методика реализации гистограммного фильтра, учитывающая следующие особенности.

Прежде всего, предполагается отказаться от единичной функции включения данных в интервал группирования, поскольку данные могут находиться вблизи границ интервала и при изменении числа интервалов оказаться в соседнем интервале. Предлагается заменить единичную функцию включения взвешенной функцией, учитывающей возможный вес «ошибочно» попавших в соседние интервалы данных:

$$u_j = \alpha_j v_{j-1} + k_j v_j + \beta_j v_{j+1},\ \alpha_j + k_j + \beta_j = 1, \qquad (1)$$

где v_j – число данных попавших в j-тый интервал группирования, $\{\alpha_j, k_j, \beta_j\}$ – весовые коэффициенты интервалов. В простейшем случае весовые коэффициенты являются постоянными величинами и могут быть выражены через коэффициент k – параметр сглаживания.

Введение весовых коэффициентов позволяет перегруппировать данные наблюдения так, чтобы уменьшилась «изрезанность» гистограммы и тем самым обеспечивалась ее сглаженность.

В работе, с использованием статистических методов, найдено выражение для коэффициента сглаживания

$$k = 1 - \frac{2}{3 + \frac{n\Delta_x^{\,4}}{2(m-1)} \int_{\gamma} \left(\frac{f''^*}{f} \right)^2 f dx}, \qquad (2)$$

где m – число интервалов, n – число данных наблюдения, Δ_x – ширина интервала группирования данных, f – предполагаемая к идентификации плотность распределения, γ – доверительный интервал.

Формула (2) позволяет сделать рад важных выводов. Во-первых, при достаточно больших объемах наблюдаемых данных коэффициент сглаживания естественным образом стремиться к единице. При малых объемах данных, когда вторая составляющая знаменателя оказывается численно незначительной, коэффициент сглаживания стремится к одной трети, что также естественным образом соответствует равномерно сглаживающему гистограммному фильтру. Во-вторых, формула (2) позволяет определять коэффициент сглаживания заранее, по имеющимся параметрам $\{n, m\}$. В-третьих, составляющая, учитывающая априорную информацию о идентифицируемом законе распределения может быть определена для класса распределений. Например, класса приближенно нормальных распределений или класса приближенно равномерных распределений.

Еще одно следствие формулы (2) заключается в том, что, задавая фиксированное небольшое расхождение коэффициента сглаживания от единицы можно получить неявную формулу связывающую число интервалов группирования данных с их объемом. Такая неявная связь двух параметров $\{n, m\}$ является теоретически обоснованной и очевидным образом зависит от формы идентифицируемого закона распределения.

Целесообразно применение полученных теоретических результатов с целью эффективной и быстрой (на малых объемах данных) 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. 26 с. 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.