УДК 519.6:519.21

# ОПТИМИЗАЦИЯ АЛГОРИТМОВ НАСТРОЙКИ КОЭФФИЦИЕНТА РАЗМЫТОСТИ ДЛЯ НЕПАРАМЕТРИЧЕСКИХ ОЦЕНОК

**Браништи В. В.**
**научный руководитель д-р техн. наук Медведев А. В.**
***Сибирский государственный аэрокосмический университет им. академика М. Ф. Решетнёва***

В различных областях науки и техники часто встречается задача оценивания неизвестной функции плотности вероятности непрерывной случайной величины. Функция плотности вероятности используется при проверке статистических гипотез, при решении различных задач классификации, распознавания образов, восстановления зависимостей и др. Одним из наиболее распространённых подходов, позволяющим по выборке значений непрерывной случайной величины восстановить её функцию плотности вероятности $f(x)$, является использование непараметрических оценок Розенблатта–Парзена. Оценка Розенблатта–Парзена функции плотности вероятности в случае одномерной случайной величины имеет вид:

$$\hat{y} = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}\Phi\left(\frac{x - x_i}{h}\right), \tag{1}$$

где $x_1, x_2, \ldots, x_n$ – независима выборка случайной величины $X$, $\Phi(z)$ – «ядро» оценки, $h$ – коэффициент размытости. Функция $\Phi(z)$ и значение $h$ существенно влияют на качество оценки и подлежат настройке.

Качество оценки (1) характеризуется усреднённой глобальной квадратичной ошибкой аппроксимации:

$$Q = \int_{-\infty}^{+\infty} M\left\{\left(\hat{f}(x) - f(x)\right)^2\right\}dx, \tag{2}$$

где $f(x)$ – истинная плотность вероятности случайной величины $X$, а математическое ожидание берётся по всей выборке. Оптимизация критерия (2), проведённая В. А. Епанечниковым, по форме ядра даёт оптимальное значение функции $\Phi(z)$ в виде усечённой параболы:

$$\Phi(z) = \begin{cases} \frac{3}{4}\left(1 - z^2\right), & |z| \le 1 \\ 0, & |z| > 1 \end{cases}, \tag{3}$$

не зависящее от истинной плотности $f(x)$ и объёма выборки $n$.

Однако полученное В. А. Епанечниковым оптимальное значение для коэффициента размытости

$$h^* = \left(\frac{1}{n}\frac{\int_{-\infty}^{+\infty}\Phi^2(z)dz}{\int_{-\infty}^{+\infty}\left(f''(x)\right)^2 dx}\right)^{1/5} \tag{4}$$

получено для $n \to \infty$ и использует вторую производную от функции плотности вероятности, следовательно, не применимо на практике.

Для настройки коэффициента размытости без использования вида истинной функции плотности вероятности можно минимизировать глобальную (неусреднённую) ошибку аппроксимации:

$$Q_0 = \int_{-\infty}^{+\infty} \left( \hat{f}(x) - f(x) \right)^2 dx \to \min_h . \quad (5)$$

Раскрыв скобки и отбросив слагаемое $\int_{-\infty}^{+\infty} f^2(x)dx$ как независящее от *h*, получим задачу оптимизации, эквивалентную задаче (5):

$$\int_{-\infty}^{+\infty} \hat{f}^2(x)dx + 2\int_{-\infty}^{+\infty} \hat{f}(x) f(x)dx \to \min_h . \quad (6)$$

Так как $\int_{-\infty}^{+\infty} \hat{f}(x) f(x)dx = M\left\{ \hat{f}(X) \right\}$, то это значение можно оценить как выборочное среднее по имеющейся выборке $x_1$, $x_2$, …, $x_n$ случайной величины *X*. Однако очевидная оценка

$$\hat{\mu}_1 = \frac{1}{n}\sum_{i=1}^{n} \hat{f}(x_i)$$

приводит к смещению:

$$M\left\{ \hat{f}(X) \right\} = \frac{1}{h} M\left\{ \Phi\left( \frac{X - Y}{h} \right) \right\};$$

$$M\{\hat{\mu}_1\} = M\left\{ \hat{f}(X) \right\} + \frac{1}{nh}\left( \Phi(0) - M\left\{ \Phi\left( \frac{X - Y}{h} \right) \right\} \right) > M\left\{ \hat{f}(X) \right\}.$$

Поэтому строится «исправленная» оценка

$$\hat{\mu}_2 = \frac{1}{n(n-1)h} \sum_{i=1}^{n} \sum_{\substack{j=1, \\ j \neq i}}^{n} \Phi\left( \frac{x_i - x_j}{h} \right), \quad (6)$$

для которой

$$M\{\hat{\mu}_2\} = M\left\{ \hat{f}(X) \right\}.$$

Используя оценку (6) можно построить функционал качества для настройки коэффициента размытости *h*, не используя вид истинной плотности вероятности:

$$W(h) = \int_{-\infty}^{+\infty} \hat{f}^2(x)dx - \frac{2}{n(n-1)h} \sum_{i=1}^{n} \sum_{\substack{j=1, \\ j \neq i}}^{n} \Phi\left( \frac{x_i - x_j}{h} \right), \quad (7)$$

который после преобразований принимает вид:

$$W(h) = \frac{1}{n^2 h^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \int_{-\infty}^{+\infty} \Phi\left( \frac{x - x_i}{h} \right) \Phi\left( \frac{x - x_j}{h} \right) dx - \frac{4}{n(n-1)h} \sum_{i=2}^{n} \sum_{j=1}^{i-1} \Phi\left( \frac{x_i - x_j}{h} \right). \quad (7')$$

В функционале (7) присутствует оператор интегрирования, что делает его вычисление затратным, а процесс минимизации – медленным. С учётом оптимальной формы ядра (3) функционал (7) принимает вид:

$$W(h) = \frac{3}{5nh} + \frac{3}{80n^2 h} \sum_{0<z_{ij}<2} \left( 2 - z_{ij} \right)^3 \left( z_{ij}^2 + 6z_{ij} + 4 \right) - \frac{3}{n(n-1)h} \sum_{0<z_{ij}<1} \left( 1 - z_{ij}^2 \right), \quad (8)$$

где $z_{ij} = \frac{x_i - x_j}{h}$, а суммирование ведётся по обоим индексам *i*, *j*, удовлетворяющим соответствующему условию.

Для минимизации функционала (8) рассмотрено 5 алгоритмов локальной оптимизации (см. табл. 1). Метод кубической интерполяции использует аналитическое

выражение для производной оптимизируемой функции, а метод Ньютона – и для производной. Из (8), с учётом того, что $z'_{ij} = -\frac{1}{h} z_{ij}$, непосредственно выводится:

$$\frac{d}{dh}W(h) = -\frac{3}{5nh^2} + \frac{3}{80n^2}\sum_{0<z_{ij}<2}\frac{d}{dh}\left(\frac{1}{h}\left(2-z_{ij}\right)^3\left(z_{ij}^2+6z_{ij}+4\right)\right) - \frac{3}{n(n-1)}\sum_{0<z_{ij}<1}\frac{d}{dh}\left(\frac{1}{h}\left(1-z_{ij}^2\right)\right) =$$

$$= \frac{3}{h^2}\left(-\frac{1}{5n} + \frac{1}{80n^2}\sum_{0<z_{ij}<2}\left(6z_{ij}^5 - 80z_{ij}^3 + 120z_{ij}^2 - 32\right) - \frac{1}{n(n-1)}\sum_{0<z_{ij}<1}\left(3z_{ij}^2-1\right)\right); \qquad (9)$$

$$\frac{d^2}{dh^2}W(h) = \frac{d}{dh}\left(\frac{3}{h^2}\left(-\frac{1}{5n} + \frac{1}{80n^2}\sum_{0<z_{ij}<2}\left(6z_{ij}^5 - 80z_{ij}^3 + 120z_{ij}^2 - 32\right) - \frac{1}{n(n-1)}\sum_{0<z_{ij}<1}\left(3z_{ij}^2-1\right)\right)\right) =$$

$$= \frac{3}{h^3}\left(\frac{2}{5n} - \frac{1}{40n^2}\sum_{0<z_{ij}<2}\left(21z_{ij}^5 - 200z_{ij}^3 + 240z_{ij}^2 - 32\right) + \frac{2}{n(n-1)}\sum_{0<z_{ij}<1}\left(6z_{ij}^2-1\right)\right). \qquad (10)$$

Для каждого алгоритма оценивалось время минимизации функционала (8) по $h$ с точностью 0,001, усреднённое по 900 запускам для различных выборок случайной величины объёма $n$ = 100. Погрешность оценивалась по правилу «3σ». Для получения результатов использовался процессор Intel® Core™ i3-2330M.

Таблица 1. Время работы алгоритмов оптимизации

| метод минимизации | время минимизации, с |
|---|---|
| метод золотого сечения | 4,11±0,05 |
| метод Пауэлла | 0,91±0,02 |
| метод кубической интерполяции | 0,99±0,04 |
| метод Ньютона | 6,0±0,4 |
| конечно-разностная аппроксимация метода Ньютона | 6,3±0,2 |

Как видно из таблицы, самыми быстродействующими оказались метод Пауэлла и метод кубической интерполяции. Метод золотого сечения показал относительно медленную сходимость. Метод Ньютона и его конечно-разностная аппроксимация показали худшие результаты, что объясняется геометрическими особенностями кривой $y = W(h)$: большой осцилляцией и малым значением первой производной в близи точки глобального минимума.

Таким образом, в приложениях при расчёте коэффициента размытости для непараметрической оценки функции плотности вероятности предлагается минимизировать критерий качества в виде (8) методом Пауэлла.