Использование метода статистических испытаний в определении мод гауссовой смеси

Н.Н.Апраушева, В.К. Де Ванса Викрамаратне, С.В.Сорокин
(Москва)

Предлагаемый алгоритм отыскания мод базируется на принципе внутренней экстремальной точки, состоящем в следующем. Если функция $f(x)$ определена и непрерывна вместе со своей производной $f'_x(x)$ на отрезке $[a,b]$, то для любой точки $x \in [a,b]$ существует отрезок $[\alpha_x, \beta_x]$ длины λ_x, $x \in [\alpha_x, \beta_x] \subset [a,b]$, в котором функция $f(x)$ или строго монотонна, или не монотонна и имеет только одну экстремальную точку $x_e \in (\alpha_x, \beta_x)$.

Исследовалась гауссова смесь с плотностью вероятности

$$f(x) = (\sqrt{2\pi}\sigma)^{-1} \sum_{i=1}^{k} \pi_i \exp(-(x-\mu_i)(2\sigma^2)^{-1}), \qquad (1)$$

где $2 \le k < \infty$, $x \in (-\infty, \infty)$, μ_i — математическое ожидание i-й компоненты, π_i — её априорная вероятность, σ^2 — дисперсия каждой компоненты,

$$\pi_i \in (0,1), \quad \sum_{i=1}^{k} \pi_i = 1.$$

Параметры распределения в (1) известны, для определённости положим

$$\mu_1 < \mu_2 < \ldots < \mu_k. \qquad (2)$$

Моды обозначим через

$$\hat{x}_1, \hat{x}_2, \ldots, \hat{x}_m, \; 1 \le m \le k.$$

Предположим, что для каждой точки μ_i, i=1,2,...,k, известна длина отрезка λ_{μ_i}, удовлетворяющего принципу внутренней экстремальной точки.

Фиксируем ту s–ю компоненту смеси, которая имеет наибольшую априорную вероятность π_s и исследуем δ-окрестность её среднего значения μ_s, где $\delta = 2^{-1}\lambda_{\mu_s}$. На отрезке $[\mu_s - \delta, \mu_s + \delta]$ моделируем n, $n \ge 10$, случайных равномерно распределённых чисел

$$\xi_1^{(1)}, \xi_2^{(1)}, \ldots, \xi_n^{(1)}$$

и определяем значения $f(\xi_i^{(1)})$, $i = 1, 2, \ldots, n$,

$$\xi_1 = \arg\max_i f(\xi_i^{(1)}).$$

1. Если

$$f(\xi_1) > f(\mu_s - \delta), \quad f(\xi_1) > f(\mu_s + \delta), \tag{3}$$

то первая мода найдена, $\widehat{x}_1 = \xi_1$.

2. Если в (3) не выполняется хотя бы одно из неравенств, например второе, то исследуем отрезок $[\mu_s + \delta, \mu_s + 2\delta]$, в котором моделируем n случайных равномерно распределённых чисел

$$\xi_1^{(2)}, \xi_2^{(2)}, \ldots, \xi_n^{(2)}.$$

Находим значения $f(\xi_i^{(2)})$, $i = 1, 2, \ldots, n$,

$$\xi_2 = \arg\max_i f(\xi_i^{(2)}).$$

Если

$$f(\xi_2) > f(\mu_s + \delta), \quad f(\xi_2) > f(\mu_s + 2\delta),$$

то $\widehat{x}_1 = \xi_2$.

Если

$$f(\mu_1 + \delta) > f(\xi_2),$$

то $\widehat{x}_1 = \mu_1 + \delta$.

Если

$$f(\mu_1 + 2\delta) > f(\xi_2),$$

то исследуем отрезок $[\mu_1 + 2\delta, \mu_1 + 3\delta]$ и т.д., пока на t–м шаге не найдём внутреннюю максимальную точку x_e отрезка $[\mu_s - \delta, \mu_s + t\delta]$, $t = 1, 2, \ldots$. Тогда $\widehat{x}_1 = x_e$. Все точки μ_i, $i \in 1, 2, \ldots, k$, лежащие в полуинтервале $[\mu_s - \delta, \mu_s + t\delta)$ исключаются из рассмотрения.

Если множество (2) не пусто, то из оставшихся компонент смеси фиксируем ту, которая имеет наибольшую априорную вероятность, и для её среднего значения повторяем всё вышеописанное исследование.

На основании теоретических результатов [1, 2] можно положить

$$\delta = 4^{-1}\sqrt{2}\sigma. \tag{4}$$

Результаты экспериментов показали, что если $\rho_{i,i+1} \geq 3$ ($\rho_{i,i+1}$ — расстояние Махаланобиса между i-й и $(i+1)$-й компонентами смеси), то принцип внутренней экстремальной точки имеет место при δ, определённом в (4).

Но возможны случаи, когда отрезок длины 2δ имеет более одной экстремальной точки. Например, функция $f(x)$ смеси с параметрами $k = 2$, $\mu_1 = 0$, $\mu_2 = 2.01$, $\sigma = 1$, $\pi_1 = \pi_2 = 0.5$, на отрезке $[0;\ 0.707]$ не имеет ни одной экстремальной точки, а на отрезке $[0.707;\ 1.414]$ имеет три внутренние экстремальные точки (две моды $\hat{x}_1 = 0.830$, $\hat{x}_2 = 1.175$ и один локальный минимум $\breve{x} = 1.005$). Наш алгоритм найдёт одну моду, например $\hat{x}_1$. Для определения другой моды исследуются отдельно отрезки $[0.707; \hat{x}_1]$ и $[\hat{x}_1; 1.414]$.

Если для смеси выполняется одно из условий унимодальности [2, 3], то исследование проводится для её математического ожидания.

Наш алгоритм определяет макромоды, но может пропустить некоторые микромоды, как и другие алгоритмы, основанные на градиентном подъёме. Так, для смеси с параметрами $k = 4$; $\sigma = 1$; $\mu_1 = 0$; $\mu_2 = 2.5$; $\mu_3 = 4.44$; $\mu_4 = 6.94$; $\pi_1 = 0.365$; $\pi_2 = 0.135$; $\pi_3 = 0.135$; $\pi_4 = 0.365$ находятся макромоды $\hat{x}_1 = 0.045$; $\hat{x}_4 = 6.895$ и пропускаются микромоды $\hat{x}_2 = 2.635$; $\hat{x}_3 = 4.305$. Для преодоления этого недостатка предлагаем использовать описанный алгоритм в сочетании с алгоритмом, основанном на методе Пикара и описанном в [1].

Литература

1. Апраушева Н.Н., Моллаверди Н. и др. О модах гауссовой смеси. М.: ВЦ РАН, 2003.
2. Aprausheva N.N., Sorokin S.V. On the unimodality of the Simple Gaussian mixture. Juor. Comput. Mathematics and Mathemat. Physics, Vol.44, No.5, 2004, 785-793.
3. Aprausheva N.N., Mollaverdi N. and Sorokin S.V. Boundaries for the Number of the Simple Gaussian Mixture Modes. Pattern recognition and image analysis, 2005. V.15. N.1. P.2004-2006.