ОБОЗРЕНИЕ
ПРИКЛАДНОЙ И ПРОМЫШЛЕННОЙ
Том 19 МАТЕМАТИКИ Выпуск 3
2012

А. К. Горшенин (Москва, ИПИ РАН)**. О модели добавления компоненты для смеси нормальных распределений.**

В работе, представленной данным докладом, рассмотрена модель добавления компоненты [1], которая является весьма удобной и информативной при проведении статистического анализа данных. Данная модель может быть использована для проверки гипотез о числе компонент смеси вероятностных распределений.

Рассмотрим сначала случай конечной масштабной смеси нормальных законов, т. е. предположим, что каждое из независимых наблюдений $\mathbf{X}_n = (X_1, X_2, \ldots, X_n)$ имеет распределение вида $G(x) = \mathbf{E}\,\Phi(Ux) = \sum_{i=1}^k p_i \Phi(x\sigma_i)$, $\sum_{i=1}^k p_i = 1$, $p_i \geqslant 0$, $\sigma_i > 0$, $i = 1, 2, \ldots, k$, где $\Phi(\cdot)$ обозначает функцию распределения стандартного нормального закона, U — дискретная случайная величина, принимающая значения σ_i с вероятностями p_i.

В этом случае модель добавления компоненты формализуется следующим образом. Пусть каждое из независимых наблюдений $\mathbf{X}_n = (X_1, X_2, \ldots, X_n)$ имеет распределение, представимое в виде $G_p(x) = (1-p)\sum_{i=1}^k p_i \Phi(x\sigma_i) + p\Phi(x\sigma)$, где все величины σ_i, p_i, $i = 1, 2, \ldots, k$, считаются известными, а σ и p являются параметрами модели, при этом $\sigma > 0$, $0 \leqslant p \leqslant 1$. Без ограничения общности для определенности будем считать, что выполнены соотношения $0 < \sigma_0 \leqslant \sigma \leqslant \sigma_1 \leqslant \sigma_2 \leqslant \cdots \leqslant \sigma_k$.

Отметим, что условие отделенности параметров от нуля является достаточно общим и означает, что рассматриваются невырожденные нормальные законы с конечными дисперсиями, поэтому величину σ_0 разумно считать известным параметром модели.

Для данной модели U_p — дискретная случайная величина, принимающая значения σ_i с вероятностями $p_i(1-p)$ и значение σ с вероятностью p. В этой ситуации расстояние Леви между смешивающими распределениями имеет вид $L(U, U_p) = p$. Тогда справедлива следующая теорема (здесь и далее $\varphi(\cdot)$ обозначает функцию плотности стандартного нормального закона).

Теорема 1. *В рамках модели добавления компоненты для конечных масштабных смесей нормальных законов при выполнении сформулированных выше условий расстояние Леви $L(U, U_p)$ между смешивающими распределениями U и U_p и расстояние Леви $L(G, G_p)$ между истинным распределением $G(x)$ и приближающей смесью $G_p(x)$ связывают неравенства*

$$\max\left\{1, \frac{\sigma_0}{\sigma_k - \sigma_0}\sqrt{2\pi e}\right\} L(G, G_p) \leqslant L(U, U_p) \leqslant \varphi^{-1/2}(\sigma_k)\left(1 + \frac{\sigma_k}{\sqrt{2\pi}}\right)^{1/2} L^{1/2}(G, G_p).$$

Теперь рассмотрим случай конечной сдвиговой смеси нормальных законов, т. е. предположим, что каждое из независимых наблюдений $\mathbf{X}_n = (X_1, X_2, \ldots, X_n)$ имеет распределение $F(x) = \mathbf{E}\,\Phi(x - V) = \sum_{i=1}^k p_i \Phi(x - a_i)$, $\sum_{i=1}^k p_i = 1$, $p_i \geqslant 0$,

$a_i \in \mathbf{R}$, $i = 1, 2, \ldots, k$, где V — дискретная случайная величина, принимающая значения a_i с вероятностями p_i. В этом случае модель добавления компоненты формализуется следующим образом. Предполагается, что каждое из независимых наблюдений $\mathbf{X}_n = (X_1, X_2, \ldots, X_n)$ имеет распределение, представимое в виде $F_p(x) = (1-p)\sum_{i=1}^{k} p_i \Phi(x - a_i) + p\Phi(x - a)$, где все величины $a_i \in \mathbf{R}$, $p_i \geqslant 0$, $i = 1, 2, \ldots, k$, считаются известными, а a и p являются параметрами модели, при этом $a \in \mathbf{R}$, $0 \leqslant p \leqslant 1$. Без ограничения общности для определенности будем считать, что выполнены соотношения $a_0 \leqslant a \leqslant a_1 \leqslant a_2 \leqslant \cdots \leqslant a_k$. Левое неравенство означает достаточно естественное для практики предположение, что рассматриваются конечные математические ожидания. Поэтому в дальнейшем считаем a_0 известным параметром модели.

В модели добавления компоненты V_p — дискретная случайная величина, принимающая значения a_i с вероятностями $p_i(1-p)$ и значение a с вероятностью p. В этой ситуации расстояние Леви между смешивающими распределениями имеет вид $L(V, V_p) = p$. Справедлива следующая теорема.

Теорема 2. *В рамках модели добавления компоненты для конечных сдвиговых смесей нормальных законов при выполнении сформулированных выше условий расстояние Леви $L(V, V_p)$ между смешивающими распределениями V и V_p и расстояние Леви $L(F, F_p)$ между истинным распределением $F(x)$ и приближающей смесью $F_p(x)$ связывают неравенства*

$$\max\left\{1, \frac{\sqrt{2\pi}}{a_k - \min\{0, a_0\}}\right\} L(F, F_p) \leqslant L(V, V_p) \leqslant \left(\frac{(1 + 1/\sqrt{2\pi}) L(F, F_p)}{\varphi(a_k + |a_k| - \min\{0, a_0\})}\right)^{1/2}.$$

Некоторые идеи доказательства приведенных теорем можно найти в [2]. В этой же работе рассмотрена еще одна удобная модель — модель расщепления компоненты. Описание особенностей ее применения, а также асимптотические результаты для данной модели можно найти в работе [3].

Работа выполнена при поддержке РФФИ, проекты 11-01-12026-офи-м и 12-07-00115.

СПИСОК ЛИТЕРАТУРЫ

1. *Бенинг В. Е., Горшенин А. К., Королев В. Ю.* Асимптотически оптимальный критерий проверки гипотез о числе компонент смеси вероятностных распределений. — Информатика и ее примен., 2011, т. 5, в. 3, с. 4–16.
2. *Горшенин А. К.* Устойчивость масштабных смесей нормальных законов относительно изменений смешивающего распределения. — Системы и средства информатики, 2012, т. 22, в. 1, с. 136–148.
3. *Горшенин А. К.* Проверка статистических гипотез в модели расщепления компоненты. — Вестник Московского Университета. Сер. 15. Вычисл. матем. и киберн., 2011, № 4, с. 26-32.