М. П. К р и в е н к о (Москва, ИПИ РАН). **Адаптивная комбинированная оценка плотности многомерного распределения.**

Оценивание плотности распределения сопровождается рядом общих для многомерного анализа и специфических для некоторых предметных областей (обработка текстов, распознавание изображений) проблем: необходимость задания множества параметров при построении оценки; существенное снижение качества оценки при увеличении размерности данных; появление вырожденности распределений при «малых» объемах обучающей выборки.

Проблемы «проклятия размерности» при использовании непараметрической оценки плотности предлагается решать с помощью перехода к главным компонентам и путем приведения данных к меньшей размерности так, чтобы сохранить специфику самих данных и по возможности улучшить выборочные свойства оценок. При переходе к главным компонентам (n-мерной переменной $\mathbf{V}$) значение плотности остается без изменений. Оценку плотности распределения $f^*(\mathbf{v})$, $\mathbf{v} = (v^{(1)}, \dots, v^{(n)})^T \in \mathbf{R}^n$, для $\mathbf{V}$ сформируем из базовой части — ядерной оценки $f_m^*(v^{(1)}, \dots, v^{(m)})$ плотности распределения первых главных компонент, и дополнительной части — параметрических оценок плотности нормального распределения $f_1^*(v^{(j)})$ остальных главных компонент: $f^*(\mathbf{v}) = f_m^*(v^{(1)}, \dots, v^{(m)}) \prod_{j=m+1}^n f_1^*(v^{(j)})$, $0 \leqslant m \leqslant n$. В работе, представленной данным сообщением,

$$f_m^*(v^{(1)}, \dots, v^{(m)}) = f_m^*(\widetilde{\mathbf{v}}) = \frac{1}{Nh^m} \sum_{i=1}^N K\biggl(\frac{\widetilde{\mathbf{v}} - \widetilde{\mathbf{v}}_i}{h}\biggr),$$

где в качестве ядра K принята плотность нормального распределения, h — параметр сглаживания, $\widetilde{\mathbf{v}}_i = (\widetilde{v}_i^{(1)}, \dots, \widetilde{v}_i^{(m)})$ суть выборочные значения, N — объем выборки.

В случае, когда $N < n$, часть диагональных элементов выборочной ковариационной матрицы $\mathbf{D}^*$ для $\mathbf{V}$ будут нулевыми. В связи с этим введем еще один параметр комбинированной оценки — критическое значение $d_0 > 0$, меньше которого не может быть значение выборочной дисперсии главных компонент.

Для подбора значения h строится вариант метода перепроверки в интерпретации [1], но применительно к многомерному случаю и в форме алгоритма, который можно использовать на практике. В этом случае значение параметра сглаживания h ищется с помощью метода перепроверки как решение задачи

$$\sum_{i=1}^N \ln\biggl(\frac{1}{h^m} \sum_{j=1,\, j\neq i}^N \exp\biggl\{-\frac{r_{ij}}{2h^2}\biggr\}\biggr) \to \max_{h>0}, \quad \text{где} \quad r_{ij} = (\widetilde{\mathbf{v}}_i - \widetilde{\mathbf{v}}_j)^T(\widetilde{\mathbf{v}}_i - \widetilde{\mathbf{v}}_j). \qquad (1)$$

Решение (1) предлагается искать итерационным путем, для эффективной реализации чего доказывается утверждение об области возможных значений точки искомого максимума. Кроме этого, аналитическим путем и с помощью метода моделирования исследуется степень влияния роста m на качество $f_m^*(\widetilde{\mathbf{v}})$.

Для исследования реальных характеристик оценок $f^*(\mathbf{v})$ рассматривается задача распознавания текста [2]. Предложенная в данном сообщении комбинированная оценка, ее экспериментальный анализ показали высокую эффективность байесовского подхода при классификации объектов, имеющих различную размерность, а также работоспособность предложенных методов оценивания элементов байесовского эмпирического классификатора.

СПИСОК ЛИТЕРАТУРЫ

1. *Duin R. P. W.* On the choice of smoothing parameters for parzen estimators of probability density functions. — IEEE Transactions on Computers, 1976, v. C-25, p. 1175–1179.

2. *Кривенко М. П.* Распознавание элементов изображения, имеющих различные размеры. — В сб.: Системы и средства информатики. В. 17. М.: ИПИ РАН, 2007, с. 30–51.