

Сравнение эффективности методов дискретизации интервала изменения значений случайной величины при синтезе непараметрической оценки плотности вероятности

А. В. ЛАПКО*, В. А. ЛАПКО**

* Институт вычислительного моделирования СО РАН, Красноярск, Россия

** Сибирский государственный аэрокосмический университет им. ак. М. Ф. Решетнева, Красноярск, Россия, e-mail: lapko@icm.krasn.ru

Исследованы аппроксимационные свойства непараметрической оценки плотности вероятности при различных методах дискретизации области определения. Оценены показатели эффективности этих методов.

Ключевые слова: непараметрическая оценка плотности вероятности, аппроксимационные свойства, методы дискретизации, нормальный закон распределения.

The approximation properties of nonparametric estimation of probability density for various applicable domain discretization methods have been studied. The indicators of the methods effectiveness have been estimated.

Key words: nonparametric estimation of probability density, approximation properties, discretization methods, normal distribution law.

Вычислительная эффективность непараметрических алгоритмов обработки информации определяется объемом статистических данных и снижается по мере его увеличения. В таких условиях целесообразно использовать принципы декомпозиции исходных статистических данных по их объему и технологию параллельных вычислений. Предложена и исследована смесь непараметрических оценок плотностей вероятности для одномерных и многомерных случайных величин [1, 2]. Показано, что она имеет значительно меньшую дисперсию по сравнению с традиционной непараметрической оценкой плотности вероятности типа Розенблатта—Парзена [3]. При этом сокращение времени вычислений сопоставимо с количеством составляющих смеси непараметрических оценок плотностей вероятности.

Полученные результаты обобщены при оценивании решающей функции в задаче распознавания образов в условиях больших выборок. Разработаны двухуровневые непараметрические системы для решения двух- и многоальтернативных задач классификации, установлены асимптотические свойства оценок уравнений разделяющих поверхностей для одномерного и многомерного случаев [4—8]. Избежать проблемы больших выборок можно, используя непараметрическую оценку плотности вероятности, синтез которой основан на сжатии исходных статистических данных [9—11]. Из условия минимума асимптотического выражения ее среднего квадратического отклонения (СКО) определена процедура оптимального выбора количества интервалов дискретизации области изменения значений случайной величины [12].

Ниже на основе анализа аппроксимационных свойств непараметрической оценки плотности вероятности проведено сравнение наиболее распространенных процедур дискретизации интервала изменения значений случайной величины с нормальным законом распределения (НЗР).

Синтез непараметрической оценки плотности вероятности. Пусть имеется выборка $V = (x^i, i = \overline{1, n})$ из n независимых значений одномерной случайной величины x с неизвестной плотностью вероятности $p(x)$. Разобьем область определения $p(x)$ на N непересекающихся интервалов длины 2β и сформируем множества случайных величин $X^j, j = \overline{1, N}$.

В качестве характеристик X^j примем частоту $\bar{P}^j$ попадания случайной величины x в j -й интервал и его центр z^j . На основе полученной информации определим массив данных $V_1 = (z^j, \bar{p}^j = \bar{P}^j / (2\beta), j = \overline{1, N})$, составленный из центров z^j введенных интервалов и соответствующих значений оценок $\bar{p}^j$ плотности вероятности. Объем N полученных данных может быть значительно меньше объема n исходной статистической информации.

В качестве приближения по эмпирическим данным V_1 искомой плотности вероятности $p(x)$ примем ее непараметрическую оценку [11]:

$$\bar{p}(x) = \frac{1}{c} \sum_{j=1}^N \bar{P}^j \Phi\left[(x - z^j)/c\right], \quad (1)$$

где ядерные функции $\Phi(u)$ удовлетворяют условиям:

$$\Phi(u) = \Phi(-u), \quad 0 \leq \Phi(u) < \infty;$$

$$\int_{-\infty}^{+\infty} \Phi(u) du = 1, \quad \int_{-\infty}^{+\infty} u^2 \Phi(u) du = 1.$$

Коэффициенты размытости с ядерных функций в (1) характеризуют область их определения.

Нетрудно убедиться, что непараметрическая оценка плотности $\bar{p}(x)$ является нормированной функцией, т. е. удовлетворяет основному свойству плотности вероятности.

Непараметрическая оценка плотности вероятности $\bar{p}(x)$ обладает свойствами асимптотической сходимости к $p(x)$ [11]. Из условия минимума асимптотического выражения СКО $\bar{p}(x)$ от $p(x)$ получим процедуру оптимального выбора количества интервалов дискретизации [12]:

$$\bar{N} = \sqrt{\Delta \|p(x)\|^2 n}, \quad (2)$$

определяемую видом восстанавливаемой плотности вероятности, областью ее определения Δ и объемом n исходных

статистических данных. Здесь $\|p(x)\|^2 = \int_{-\infty}^{+\infty} p^2(x) dx$. Получен-

ная закономерность является объективной, так как не зависит от вида ядерных функций непараметрической оценки плотности вероятности (1).

Анализ аппроксимационных свойств непараметрической оценки плотности вероятности. Исследуем зависимость аппроксимационных свойств непараметрической оценки плотности вероятности (1) от методов дискретизации области изменения значений одномерной случайной величины x и объема n исходных статистических данных $V = (x^i, i = \overline{1, n})$.

Восстановим плотность вероятности случайной величины с НЗР:

$$p(x) = (2\pi)^{-1/2} \exp(-x^2/2).$$

Для выбора количества интервалов дискретизации области изменения значений случайной величины используем (2), а также формулы:

$$\text{Старджесса } N = \log_2 n + 1;$$

$$\text{Брукса и Каррузера } N = 5 \lg n;$$

$$\text{Хайнкольда и Гаеде } N = \sqrt{n};$$

$$\text{И. У. Алексеевой } N = \frac{4}{\chi} \lg \frac{n}{10},$$

где $\chi = 1/\sqrt{3}$ — контрэксцесс НЗР.

При использовании (2) имеем $\sqrt{\Delta \|p(x)\|^2} = 1,3$.

Синтез непараметрической оценки плотности вероятности (1) осуществим на основе ядерных функций В. А. Епанечникова [13]:

$$\Phi(u) = \begin{cases} 3/(4\sqrt{5}) - 3u^2/(20\sqrt{5}) & \forall |u| < \sqrt{5}, \\ 0 & \forall |u| \geq \sqrt{5}. \end{cases}$$

В данных условиях качество восстановления $p(x)$ определим статистическими оценками критериев:

$$W_1 = M|p(x) - \bar{p}(x)|; \quad W_2 = M(p(x) - \bar{p}(x))^2;$$

$$W_3 = M(p(x) - \bar{p}(x))^2 - (M(p(x) - \bar{p}(x)))^2,$$

которые являются смещением, СКО и дисперсией регрессионной оценки плотности вероятности $\bar{p}(x)$, соответственно. Здесь M — математическое ожидание.

Для оценивания W_1, W_2 в виде средних используем контрольную выборку значений случайной величины x объемом $n_k = 200$. По полученным значениям $\tilde{W}_1, \tilde{W}_2$ вычислим оценку $\tilde{W}_3$.

При одних и тех же объемах n статистических данных V многократно ($m = 100$) определим значения $\tilde{W}_1^j, \tilde{W}_2^j, \tilde{W}_3^j$,

$j = \overline{1, m}$, которые в дальнейшем усредним $\bar{W}_v = \frac{1}{m} \sum_{j=1}^m \tilde{W}_v^j$,

$v = \overline{1, 3}$ и сопоставим со значением n (см. рисунок).

Количественные оценки показателей эффективности применения сравниваемых формул дискретизации области изменения значений случайной величины с НЗР приведены в табл. 1—3.

Таблица 1

Статистические оценки смещения $\bar{W}_1$ регрессионной оценки плотности вероятности (1)

n	Оценка $\bar{W}_1$ по формулам дискретизации				
	Старджесса	Брукса и Каррузера	Хайнкольда и Гаеде	И. У. Алексеевой	Оптимальный выбор количества интервалов
50	0,01262	0,01165	0,01262	0,01802	0,01062
100	0,00759	0,00643	0,00643	0,00844	0,00522
150	0,00586	0,00463	0,00426	0,00586	0,00330
200	0,00423	0,00359	0,00308	0,00423	0,00247
250	0,00383	0,00329	0,00235	0,00353	0,00199
300	0,00317	0,00293	0,00187	0,00309	0,00146
350	0,00301	0,00241	0,00163	0,00256	0,00130
400	0,00248	0,00212	0,00130	0,00242	0,00101
450	0,00224	0,00196	0,00115	0,00200	0,00089
500	0,00204	0,00188	0,00103	0,00196	0,00083

Таблица 2

Статистические оценки СКО $\bar{W}_2$ регрессионной оценки плотности вероятности (1)

n	Оценка $\bar{W}_2$ по формулам дискретизации				
	Старджесса	Брукса и Каррузера	Хайнкольда и Гаеде	И. У. Алексеевой	Оптимальный выбор количества интервалов
50	0,00310	0,00298	0,00310	0,00430	0,00271
100	0,00241	0,00201	0,00201	0,00271	0,00157
150	0,00240	0,00177	0,00162	0,00240	0,00130
200	0,00188	0,00150	0,00128	0,00188	0,00101
250	0,00196	0,00158	0,00112	0,00173	0,00094
300	0,00192	0,00158	0,00100	0,00170	0,00081
350	0,00183	0,00131	0,00082	0,00138	0,00064
400	0,00153	0,00123	0,00072	0,00144	0,00057
450	0,00151	0,00120	0,00068	0,00126	0,00052
500	0,00151	0,00128	0,00066	0,00133	0,00051

Таблица 3

Статистические оценки дисперсии $\bar{W}_3$ регрессионной оценки плотности вероятности (1)

n	Оценка $\bar{W}_3$ по формулам дискретизации				
	Старджесса	Брукса и Каррузера	Хайнкольда и Гаеде	И. У. Алексеевой	Оптимальный выбор количества интервалов
50	0,00289	0,00281	0,00289	0,00393	0,00257
100	0,00234	0,00195	0,00195	0,00263	0,00153
150	0,00236	0,00174	0,00159	0,00236	0,00128
200	0,00186	0,00148	0,00127	0,00186	0,00100
250	0,00194	0,00156	0,00111	0,00172	0,00093
300	0,00191	0,00157	0,00100	0,00169	0,00080
350	0,00182	0,00131	0,00082	0,00137	0,00063
400	0,00152	0,00122	0,00072	0,00143	0,00057
450	0,00150	0,00120	0,00068	0,00126	0,00052
500	0,00150	0,00128	0,00066	0,00132	0,00051

С ростом n применение исследуемых методов дискретизации интервала изменений случайной величины приводит к уменьшению $\bar{W}_1$, $\bar{W}_2$, $\bar{W}_3$ (см. рисунок, табл. 1—3). Наблюдаемое улучшение аппроксимационных свойств $\bar{p}(x)$ можно объяснить увеличением объема N статистической выборки V_1 , используемой при синтезе непараметрической оценки плотности вероятности. Данный факт согласуется с условиями ее асимптотической сходимости [11, 12].

Установлено, что применение (2) при выборе количества N интервалов дискретизации предпочтительнее по сравнению с другими формулами. Данный вывод вполне ожидаем, так как (2) оптимальна в смысле минимума асимптотического выражения СКО непараметрической оценки плотности вероятности. Показатели эффективности использования формул Брукса и Каррузера, Хайнкольда и Гаеде при синтезе $\bar{p}(x)$ сопоставимы при

$n < 200$ (см. рисунок). При $n > 200$ применение формул Старджесса, Брукса и Каррузера и И. У. Алексеевой позволяет получить близкие значения показателей $\bar{W}_1$, $\bar{W}_2$, $\bar{W}_3$, определяющих аппроксимационные свойства непараметрической оценки плотности вероятности (см. табл. 1—3). При $n < 100$ использование формулы И. У. Алексеевой для выбора количества интервалов дискретизации N нецелесообразно, а результативность формул Старджесса, Брукса и Каррузера и Хайнкольда и Гаеде сопоставима (см. табл. 1—3).

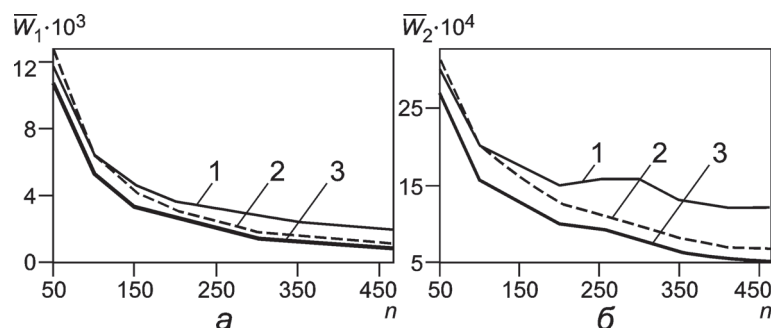
Выводы. Методика синтеза непараметрической оценки плотности вероятности (1) предполагает использование процедуры дискретизации области изменения значений случайной величины. Поэтому появляется возможность оценить эффективность различных формул дискретизации и на этой основе осуществить их анализ.

При восстановлении плотности вероятности случайной величины с НЗР целесообразно использовать формулы оптимального выбора количества интервалов и Хайнкольда и Гаеде, менее предпочтительны формулы Старджесса и И. У. Алексеевой. Формула дискретизации Брукса и Каррузера по своей эффективности занимает промежуточное положение.

Полученные результаты имеют важное значение при проверке гипотез о распределениях случайных величин, для построения алгоритма обработки данных в разных устройствах при работе с изображениями.

Литература

1. Лапко А. В., Лапко В. А., Егорочкин И. А. Непараметрические оценки смеси плотностей вероятности и их применение в задаче распознавания образов // Системы управления и информационные технологии. 2009. Т. 35. № 1. С. 60—64.
2. Лапко А. В., Лапко В. А. Синтез структуры смеси непараметрических оценок плотности вероятности многомерной случайной величины // Системы управления и информационные технологии. 2011. Т. 43. № 1. С. 12—15.
3. Parzen E. On estimation of a probability density function and mode // Ann. math. statistic. 1962. Т. 33. N. 3. P. 1065—1076.



Зависимости статистических оценок смещения $\bar{W}_1$ (а) и среднеквадратического отклонения $\bar{W}_2$ (б) непараметрической оценки плотности вероятности $\bar{p}(x)$ от объема n исходных данных V ; кривые 1—3 соответствуют формулам дискретизации Брукса и Каррузера, Хайнкольда и Гаеде, (2)

4. Лапко А. В., Лапко В. А. Коллектив непараметрических решающих функций в двальтернативной задаче распознавания образов // Системы управления и информационные технологии. 2009. Т. 37. № 3. С. 156—160.

5. Лапко А. В., Лапко В. А. Разработка и исследование двухуровневых непараметрических систем классификации // Автометрия. 2010. Т. 46. № 1. С. 70—78.

6. Lapko A. V., Lapko V. A. Development and Investigation of Two-Level Non-Parametric Estimators // Opt., Instrument. Data Proces. 2010. Т. 46. N. 1. P. 56—63.

7. Лапко А. В., Лапко В. А. Синтез структуры семейства непараметрических решающих функций в задаче распознавания образов // Автометрия. 2011. Т. 47. № 4. С. 76—82.

8. Lapko A. V., Lapko V. A. Synthesis of the Structure of a Family of Nonparametric Decision Functions in the Pattern Recognition Problem // Opt., Instrument. Data Proces. 2011. Т. 47. N. 4. P. 383—387.

9. Лапко А. В., Лапко В. А. Непараметрические методики анализа множеств случайных величин // Автометрия. 2003. Т. 39. № 1. С. 54—61.

10. Lapko A. V., Lapko V. A. Non-parametric Analysis Techniques of Random Values sets // Opt., Instrument. Data Proces. 2003. Т. 39. N. 1. P. 44—50.

11. Лапко А. В., Лапко В. А. Регрессионная оценка плотности вероятности и ее свойства // Системы управления и информационные технологии. 2012. Т. 49. № 3. С. 152—156.

12. Лапко А. В., Лапко В. А. Оптимальный выбор количества интервалов дискретизации области изменения одномерной случайной величины при оценивании плотности вероятности // Измерительная техника. 2013. № 7. С. 24—27; Lapko A. V., Lapko V. A. Optimal selection of the number of sampling intervals in domain of variation of a one-dimensional random variable in estimation of the probability density // Measurement Techniques. 2013. V. 56. N. 7. P. 763—767.

13. Епанечников В. А. Непараметрическая оценка многомерной плотности вероятности // Теория вероятности и ее применения. 1969. Т. 14. № 1. С. 156—161.

Дата принятия 30.12.2013 г.

НАНОМЕТРОЛОГИЯ

621.00.56

Разработка экспериментально-аналитической методики определения температурных деформаций концевых мер длины в нанометровом диапазоне

О. С. БАШЕВСКАЯ*, С. В. БУШУЕВ*, Ю. В. ПОДУРАЕВ*, М. Г. КОВАЛЬСКИЙ**, Г. Б. КАЙНЕР**, Е. В. РОМАШ*, Е. А. МЕЛЬНИЧЕНКО*

* Московский государственный технологический университет «СТАНКИН», Москва, Россия, e-mail: bashevskaya@mail.ru

** ОАО «НИИИзмерения», Москва, Россия, e-mail: kovalskiy@micron.ru

Рассмотрено влияние температурных деформаций концевых мер длины на процесс контроля прецизионных деталей. Предложена и экспериментально апробирована экспериментально-аналитическая методика определения температурных деформаций концевых мер длины, используемая при проведении линейных измерений в нанометровом диапазоне.

Ключевые слова: температурные деформации, концевые меры длины, нанометровый диапазон.

The influence of thermal strains of end standards of meter on the precision elements control process is considered. The experimental-analytical procedure of determination of end standards of meter temperature deformations used at linear nanoscale measurements is suggested and tested experimentally.

Key words: temperature deformations, end standards of meter, nanometer range.

При контроле прецизионных деталей в нанометровом диапазоне методом сравнения, а также для поверки и калибровки средств измерений (СИ) геометрических параметров широко используют концевые меры длины (КМД). При воспроизведении длины КМД обеспечивают идентичность

условий и состояний этого процесса, идентичность внутренних напряжений и нанодеформаций [1, 2]. Развитие нанотехнологий доказывает необходимость продолжения исследований, направленных на обеспечение единства измерений в нанометровом диапазоне 1—100 нм [3, 4].