УДК 519.7

ЗАВИСИМОСТЬ СВОЙСТВ РЕГРЕССИОННОЙ ОЦЕНКИ ПЛОТНОСТИ ВЕРОЯТНОСТИ ОТ ОСОБЕННОСТЕЙ МЕТОДИКИ ЕЁ СИНТЕЗА*

Д. В. Борисов[1], А. В. Лапко[1, 2], В. А. Лапко[1, 2]

[1]Сибирский государственный аэрокосмический университет имени академика М. Ф. Решетнева
Российская Федерация, 660014, Красноярск, просп. им. газ. «Красноярский рабочий», 31
E-mail: valapko@yandex.ru
[2]Институт вычислительного моделирования СО РАН
Российская Федерация, 660036, г. Красноярск, Академгородок, 50, стр. 44. E-mail: lapko@icm.krasn.ru

Исследуются аппроксимационные свойства регрессионной оценки плотности вероятности. Синтез оценки основывается на декомпозиции исходных статистических данных и анализе вероятностных характеристик получаемых множеств случайных величин. Устанавливается зависимость свойств регрессионной оценки плотности вероятности от методов дискретизации интервала значений случайной величины. Из условия минимума асимптотического выражения среднеквадратического отклонения определена процедура оптимального выбора количества интервалов дискретизации. Полученная формула зависит от вида восстанавливаемой плотности вероятности и объёма априорных данных. Результаты исследований имеют важное значение при решении задач проверки гипотез о распределениях случайных величин и доверительного оценивания плотности вероятности.

Ключевые слова: плотность вероятности, регрессионная оценка, аппроксимационные свойства, методы дискретизации.

DEPENDENCE OF THE REGRESSION ESTIMATOR PROPERTIES OF A PROBABILITY DENSITY ON SINGULARITIES OF ITS SYNTHESIS TECHNIQUE

D. V. Borisov[1], A. V. Lapko[1, 2], V. A. Lapko[1, 2]

[1]Siberian State Aerospace University named after academician M. F. Reshetnev
31, Krasnoyarsky Rabochy Av., Krasnoyarsk, 660014, Russian Federation
E-mail: valapko@yandex.ru
[2]Institute of Computational Modeling, Siberian Branch of RAS
50, Akademgorodok, Krasnoyarsk, 660036, Russian Federation
E-mail: lapko@icm.krasn.ru

Approximating properties of the regression estimator of a probability density are investigated. Estimation synthesis is based on decomposition of initial statistical data and the analysis of probabilistic characteristics of received sets of random variables. Dependence of the regression estimator properties of a probability density on methods of digitization of an interval of values of a random variable is established. A deviation mean square procedure of an optimum choice of an amount of intervals of digitization is defined from a condition of a minimum of asymptotic expression. The received formula depends on an aspect of a restored probability density and volume of a priori data. Outcomes of researches are important to the solution of the problems of a hypothesis test about distributions of random variables and a confidential estimation of a probability density.

Keywords: probability density, regression estimator, approximating properties, digitization methods.

Непараметрические оценки плотности вероятности типа Розенблатта–Парзена широко используются при синтезе алгоритмов обработки информации и принятии решений в условиях априорной неопределённости [1–11]. Однако их вычислительная эффективность во многом определяется объёмом статистических данных и снижается по мере его увеличения.

В данных условиях целесообразно использовать принципы декомпозиции исходных статистических данных по их объёму и технологию параллельных вычислений. С этих позиций предложена и исследована смесь непараметрических оценок плотностей вероятности для одномерных и многомерных случайных величин [4; 12; 13].

Перспективное направление решения проблем больших выборок связано с декомпозицией исходных статистических данных и последующим анализом вероятностных характеристик получаемых множеств случайных величин [14; 15].

Пусть имеется выборка $V = \left(x^i , i = \overline{1, n} \right)$ из n независимых значений одномерной случайной величины x с неизвестной плотностью вероятности $p(x)$. Разобьем область определения $p(x)$ на N непересекающихся

* Работа выполнена в рамках базовой части государственного задания Минобрнауки РФ (СибГАУ № Б121/14).

интервалов длиной 2β и сформируем множества случайных величин X^j, $j=\overline{1,N}$. В качестве характеристик X^j примем частоту $\overline{P}^j$ попадания случайной величины x в j-й интервал и его центр z^j. На основе полученной информации определим массив данных $V_1=\left(z^j,\overline{P}^j/(2\beta),j=\overline{1,N}\right)$, составленный из центров z^j введенных интервалов и соответствующих им значений оценок плотности вероятности. Объём N полученных данных V_1 может быть значительно меньше объёма n исходной статистической информации V.

В качестве приближения по эмпирическим данным V_1 искомой плотности вероятности $p(x)$ примем статистику [15]

$$\overline{p}(x)=c^{-1}\sum_{j=1}^{N}\overline{P}^j\Phi\left(\frac{x-z^j}{c}\right), \qquad (1)$$

в которой ядерные функции $\Phi(u)$ являются положительными, симметричными и нормированными [16]. Коэффициенты размытости c ядерных функций характеризуют область их определения.

В работе исследуется зависимость аппроксимационных свойств регрессионной оценки плотности вероятности (1) от известных методов дискретизации области изменения значений случайной величины.

Выбор оптимального количества интервалов дискретизации области значений случайной величины. В работе [17] исследованы свойства среднеквадратического отклонения

$$M\int_{-\infty}^{+\infty}\left(p(x)-\overline{p}(x)\right)^2dx$$

оценки $\overline{p}(x)$ (1) от восстанавливаемой плотности вероятности $p(x)$ при больших объёмах исходных статистических данных, где M – знак математического ожидания. При оптимальных значениях коэффициентов размытости получено его асимптотическое выражение

$$W_2(N)=\left(\frac{\left(\left\|\Phi(u)\right\|^2\right)^4\left\|p^{(2)}(x)\right\|^2}{2N^4}\right)^{\frac{1}{5}}\times \\ \times\left(\frac{3}{2}\left(\Delta\left\|p(x)\right\|^2\right)^{4/5}+\frac{N^2}{n\left(\Delta\left\|p(x)\right\|^2\right)^{1/5}}\right), \qquad (2)$$

где

$$\left\|\Phi(u)\right\|^2=\int_{-\infty}^{+\infty}\Phi^2(u)du,\ \left\|p(x)\right\|^2=\int_{-\infty}^{+\infty}p^2(x)dx,$$

$$\left\|p^{(2)}(x)\right\|^2=\int_{-\infty}^{+\infty}\left(p^{(2)}(x)\right)^2dx;$$

Δ – длина интервала изменения значений случайной величины.

Из условия минимума асимптотического выражения среднеквадратического отклонения $\overline{p}(x)$ от $p(x)$ получена процедура оптимального выбора количества интервалов дискретизации [17; 18]

$$\overline{N}=\sqrt{\Delta\left\|p(x)\right\|^2n}\ , \qquad (3)$$

которая определяется видом восстанавливаемой плотности вероятности, значением Δ и объёмом n исходных статистических данных. Полученная закономерность является объективной, так как не зависит от вида ядерных функций оценки плотности вероятности (1).

Исследование аппроксимационной регрессионной оценки плотности вероятности. Будем восстанавливать плотность вероятности случайной величины с нормальным законом распределения

$$p(x)=\frac{1}{\sqrt{2\pi}}\exp\left(\frac{-x^2}{2}\right).$$

Для выбора количества интервалов дискретизации области изменения значений случайной величины используется выражение (3), а также следующие формулы:

– Хайнкольда и Гаеде

$$N=\sqrt{n}\ ; \qquad (4)$$

– Брукса и Каррузера

$$\mathrm{N}=5\lg n\ ; \qquad (5)$$

– Старджесса

$$\mathrm{N}=\log_2 n+1\ . \qquad (6)$$

Синтез непараметрической оценки плотности вероятности (1) осуществлялся на основе ядерных функций В. А. Епанечникова [16]

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

В данных условиях выражение (2) запишется в виде

$$\overline{W}_2=\frac{3}{10}\left(\frac{2}{15N^4}\right)^{\frac{1}{5}}\left(\frac{9}{2\sqrt{\pi}}+\frac{N^2}{n}\right).$$

При увеличении объёма n исходных статистических данных применение исследуемых методов дискретизации интервала изменения значений случайной величины приводит к уменьшению значений $\overline{W}_2$ (см. рисунок). Наблюдаемое улучшение аппроксимационных свойств $\overline{p}(x)$ объясняется увеличением объёма N массива данных V_1, используемого при построении регрессионной оценки плотности вероятности (см. таблицу). Данный факт согласуется с условиями её асимптотической сходимости [15].

Зависимость количества интервалов N от значений n и используемых формул дискретизации

n	Формулы дискретизации			
	(3)	(4)	(5)	(6)
50	9	7	8	7
100	13	10	10	8
150	16	12	11	8
200	18	14	12	9
250	21	16	12	9
300	23	17	12	9

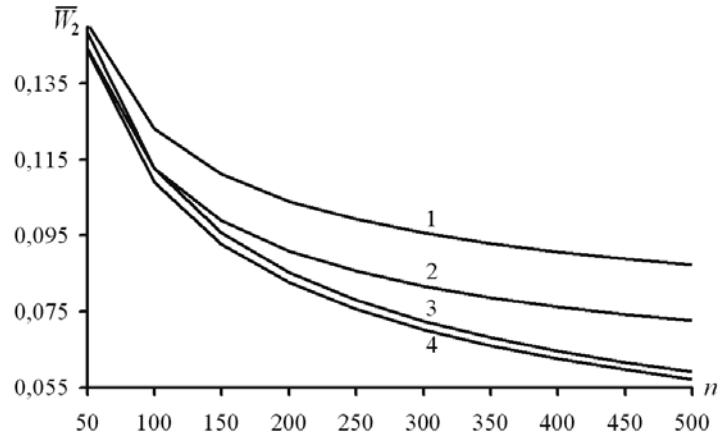
Окончание таблицы

n	Формулы дискретизации			
	(3)	(4)	(5)	(6)
350	24	19	13	9
400	26	20	13	10
450	28	21	13	10
500	29	22	13	10

Применение формулы (3) при выборе количества *N* интервалов дискретизации является более предпочтительным по сравнению с другими, так как она получена на основе минимизации асимптотического выражения среднеквадратического отклонения (2).

Зависимости $\overline{W}_2$ от объёма *n* исходных данных при использовании формул (3), (4) являются близкими. Им свойственны сопоставимые значения количества *N* интервалов дискретизации области изменения случайной величины (см. таблицу). При малых $n < 100$ количество *N* интервалов дискретизации, которые определяются формулами (3)–(6), и соответствующие им значения $\overline{W}_2$ отличаются незначительно.

При восстановлении плотности вероятности с нормальным законом распределения целесообразно использовать формулы (3), (4). Менее предпочтительными являются формулы (5), (6). Полученные выводы согласуются с результатами исследований асимптотических свойств регрессионной оценки плотности вероятности.



Зависимость среднеквадратического отклонения $\overline{W}_2(N)$
от объёма *n* значений случайной величины с нормальным законом распределения:
кривые 1, 2, 3, 4 соответствуют значениям *N*, вычисленным по формулам (3)–(6)

Библиографические ссылки

1. Лапко А. В., Лапко В. А. Гибридные модели стохастических зависимостей // Автометрия. 2002. № 5. С. 38–48.

2. Лапко В. А., Капустин А. Н. Синтез нелинейных непараметрических коллективов решающих правил в задачах распознавания образов // Автометрия. 2006. Т. 42, № 6. С. 26–33.

3. Лапко А. В., Лапко В. А. Анализ непараметрических алгоритмов распознавания образов в условиях пропуска данных // Автометрия. 2008. Т. 44, № 3. С. 65–74.

4. 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. № 1 (35). С. 60–64.

5. Лапко А. В., Лапко В. А. Коллектив непараметрических решающих функций в двуальтернативной задаче распознавания образов // Системы управления и информационные технологии. 2009. № 3.1 (37). С. 156–160.

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

7. Лапко А. В., Лапко В. А. Асимптотические свойства многомерной непараметрической оценки уравнения разделяющей поверхности в двуальтернативной задаче распознавания образов // Системы управления и информационные технологии. 2010. № 1 (39). С. 16–19.

8. Лапко А. В., Лапко В. А. Непараметрическая оценка уравнения разделяющей поверхности в условиях больших выборок и её свойства // Системы управления и информационные технологии. 2010. № 1.2 (39). С. 300–304.

9. Лапко А. В., Лапко В. А. Применение непараметрического алгоритма распознавания образов в задаче проверки гипотезы о распределениях случайных величин // Системы управления и информационные технологии. 2010. № 3 (41). С. 8–11.

10. Лапко А. В., Лапко В. А. Непараметрические алгоритмы распознавания образов в задаче проверки статистической гипотезы о тождественности двух законов распределения случайных величин // Автометрия. 2010. Т. 46, № 6. С. 47–53.

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

12. Лапко А. В., Лапко В. А. Синтез структуры смеси непараметрических оценок плотности вероятности многомерной случайной величины // Системы управления и информационные технологии. 2011. № 1 (43). С. 12–15.

13. Лапко А. В., Лапко В. А. Анализ свойств непараметрических оценок смеси плотностей вероятности при различных условиях распределения статистических данных // Информатика и системы управления. 2013. № 1 (35). С. 119–126.

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

15. Лапко А. В., Лапко В. А. Регрессионная оценка плотности вероятности и ее свойства // Системы управления и информационные технологии. 2012. № 3 (49). С. 152–156.

16. Епанечников В. А. Непараметрическая оценка многомерной плотности вероятности // Теория вероятности и ее применения. 1969. Т. 14. Вып. 1. С. 156–161.

17. 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.

18. 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. Vol. 56, no. 7. P. 24–27 (DOI: 10.1007/s11018-013-0279-x).

References

1. Lapko A. V., Lapko V. A. *Avtometriya*. 2002, no. 5, p. 38–48.

2. Lapko V. A., Kapustin A. N. *Avtometriya*. 2006, vol. 42, no. 6, p. 26–33.

3. Lapko A. V., Lapko V. A. *Avtometriya*. 2008, vol. 44, no. 3, p. 65–74.

4. Lapko A. V., Lapko V. A., Egorochkin I. A. *Sistemy upravlenija i informacionnye tehnologii*. 2009, no. 1 (35), p. 60–64.

5. Lapko A. V., Lapko V. A. *Sistemy upravlenija i informacionnye tehnologii*. 2009, no. 3.1 (37), p. 156–160.

6. Lapko A. V., Lapko V. A. *Avtometriya*. 2010, vol. 46, no. 1, p. 70–78.

7. Lapko A. V., Lapko V. A. *Sistemy upravlenija i informacionnye tehnologii*. 2010, no. 1 (39), p. 16–19.

8. Lapko A. V., Lapko V. A. *Sistemy upravlenija i informacionnye tehnologii*. 2010, no. 1.2 (39), p. 300–304.

9. Lapko A. V., Lapko V. A. *Sistemy upravlenija i informacionnye tehnologii*. 2010, no. 3 (41), p. 8–11.

10. Lapko A. V., Lapko V. A. *Avtometriya*. 2010, vol. 46, no. 6, p. 47–53.

11. Lapko A. V., Lapko V. A. *Avtometriya*. 2011, vol. 47, no. 4, p. 76–82.

12. Lapko A. V., Lapko V. A. *Sistemy upravlenija i informacionnye tehnologiju*. 2011, no. 1 (43), p. 12–15.

13. Lapko A. V., Lapko V. A. *Informatika i sistemy upravlenija*. 2013, no. 1 (35), p. 119–126.

14. Lapko A. V., Lapko V. A. *Avtometriya*. 2003, vol. 39, no. 1, p. 54–61.

15. Lapko A. V., Lapko V. A. *Sistemy upravlenija i informacionnye tehnologii*. 2012, no. 3 (49), p. 152–156.

16. Epanechnikov V. A. *Teorija verojatnosti i ee primenenija*. 1969, vol. 14, no. 1, p. 156–161.

17. Lapko A. V., Lapko V. A. *Izmeritel'naja tehnika*. 2013, no. 7, p. 24–27.

18. 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 (2013) *Measurement Techniques*, 56 (7), p. 24–27. doi: 10.1007/s11018-013-0279-x.

УДК 004.93

INFORMATIVE ATTRIBUTE SELECTION WITH HYBRID SELF-ADJUSTED EVOLUTIONARY OPTIMIZATION ALGORITHM*

S. S. Volkova

Siberian State Aerospace University named after academician M. F. Reshetnev
31, Krasnoyarskiy Rabochiy Av., Krasnoyarsk, 660014, Russian Federation
E-mail: Svetlana.volkova.mail@yandex.ru

An informative attribute selection problem is considered. The problem is solved with the hybrid self-adjusted evolutionary algorithm. The algorithm is used as an optimization method of bandwidth parameters in kernel regression. The algorithm is experimented on the test function with various dimensions. Reliability depends on the dimension function which is also presented. The results of the hybrid self-adjusted algorithm are presented too.

Keywords: informative attribute selection, kernel regression, genetic algorithm, hybrid self-adjusted genetic algorithm.

* The study was supported by The Ministry of education and science of Russian Federation, project № 14.B37.21.1521.
The second International Workshop on Mathematical Models and its Applications (IWMMA 2013).