УДК 519.24

© 2015 г. **А.В. Лапко**, д-р техн. наук,
В.А. Лапко, д-р техн. наук
(Институт вычислительного моделирования СО РАН, Красноярск,
Сибирский государственный аэрокосмический университет имени
академика М.Ф. Решетнева, Красноярск)

АНАЛИЗ ЭФФЕКТИВНОСТИ МЕТОДОВ ДИСКРЕТИЗАЦИИ ИНТЕРВАЛА ИЗМЕРЕНИЙ СЛУЧАЙНОЙ ВЕЛИЧИНЫ ПРИ ОЦЕНИВАНИИ ПЛОТНОСТИ ВЕРОЯТНОСТИ*

Определяется количественная зависимость аппроксимационных свойств непараметрической оценки плотности вероятности от методов дискретизации интервала измерений случайной величины. Исследуется ее особенность от объема исходных статистических данных и вида восстанавливаемой плотности вероятности.

Ключевые слова: обработка измерений, плотность вероятности, регрессионная оценка, методы дискретизации.

Введение

К одной из актуальных задач теории математической обработки измерений, требующих своего решения, относится эффективный выбор количества интервалов дискретизации области изменения значений случайных величин. Данная проблема возникает, например, при оценивании плотности вероятности и проверке гипотез о распределениях случайных величин с использованием критерия Пирсона [1 – 3]. Перспективное направление ее решения связано с исследованием аппроксимационных свойств регрессионной оценки плотности вероятности, синтез которой основан на декомпозиции исходных статистических данных и последующем анализе вероятностных характеристик получаемых множеств случайных величин [4, 5].

В данной работе на основе анализа свойств регрессионной оценки плотности вероятности исследуется эффективность процедур дискретизации интервала измерений случайной величины для наиболее распространенных законов распределения.

* Работа выполнена в рамках базовой части государственного задания Министерства образования и науки РФ высшим учебным заведениям на 2014 – 2016 гг. (СибГАУ № Б121/14).

Регрессионная оценка плотности вероятности

Пусть имеется выборка $V = \left(x^i, i = \overline{1, n}\right)$ из n независимых измерений одномерной случайной величины x с неизвестной плотностью вероятности $p(x)$.

Разобьем область определения $p(x)$ на N непересекающихся 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 $V_1 = \left(z^j, \overline{p}^j = \overline{P}^j/(2\beta), j = \overline{1, N}\right)$, составленный из центров z^j введенных интервалов и соответствующих им значений оценок $\overline{p}^j$ 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 регрессионную оценку [4]

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

где ядерные функции $\Phi(u)$ удовлетворяют свойствам положительности, симметричности относительно значений z^j и нормированности [6]. Коэффициенты размытости c ядерных функций в выражении (1) характеризуют область их определения.

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

Регрессионная оценка плотности вероятности $\overline{p}(x)$ обладает свойствами асимптотической сходимости к $p(x)$ [7]. Из условия минимума асимптотического выражения среднеквадратического отклонения $\overline{p}(x)$ от $p(x)$ получена процедура оптимального выбора количества интервалов дискретизации [7, 8]

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

которая зависит от вида восстанавливаемой плотности вероятности, области ее определения Δ и объема n исходных статистических данных. Здесь выражение $\left\| p(x) \right\|^2$ соответствует интегралу от $p^2(x)$.

Сравнение эффективности методов дискретизации

В качестве основы при формировании критерия сравнения методов дискретизации будем использовать оценку [7]

$$W_2 = \frac{2\Delta}{Nc} \left\| p(x) \right\|^2 \left\| \Phi(u) \right\|^2 + \frac{c^4}{4} \left\| p^{(2)}(x) \right\|^2 \qquad (3)$$

среднеквадратического отклонения $M\int\limits_{-\infty}^{+\infty}(\overline{p}(x)-p(x))^2\,dx$ регрессионной оценки $\overline{p}(x)$ от плотности вероятности $p(x)$. Здесь M – знак математического ожидания, $\|\Phi(u)\|^2=\int\limits_{-\infty}^{+\infty}\Phi^2(u)\,du$, $\left\|p^{(2)}(x)\right\|^2=\int\limits_{-\infty}^{+\infty}\left(p^{(2)}(x)\right)^2dx$.

Определим значение (3) при оптимальном коэффициенте размытости $\overline{c}$ ядерных функций статистики (1). Из условия минимума W_2 по c получим

$$\overline{c}=\left(\frac{2\Delta\,\|\Phi(u)\|^2\|p(x)\|^2}{N\left\|p^{(2)}(x)\right\|^2}\right)^{\frac{1}{5}}.$$

Тогда при $c=\overline{c}$ выражение (3) перепишется в виде

$$\overline{W}_2=\frac{5}{4}\left[\left(\frac{2\Delta\,\|\Phi(u)\|^2\|p(x)\|^2}{N}\right)^4\left\|p^{(2)}(x)\right\|^2\right]^{\frac{1}{5}}. \qquad (4)$$

Исследуем влияние методов дискретизации интервала измерений случайной величины на аппроксимационные свойства регрессионной оценки плотности вероятности. Для этого определим отношения выражений типа (4), в которых количество N интервалов дискретизации вычисляется в соответствии с процедурой (2) и формулами:

Хайнкольда и Гаеде [9]

$$N=\sqrt{n}\,, \qquad (5)$$

Брукса и Каррузера [10]

$$N=5\,\lg n\,, \qquad (6)$$

Старджесса [11]

$$N=\log_2 n+1. \qquad (7)$$

Обозначим через $\overline{W}_2(2)$ значение выражения (4) полученное, например, при использовании формулы дискретизации (2).

Тогда нетрудно показать справедливость следующих отношений:

$$R(2,5)=\frac{\overline{W}_2(2)}{\overline{W}_2(5)}=\left(\frac{1}{\sqrt{\Delta\,\|p(x)\|^2}}\right)^{\frac{4}{5}},\quad R(2,6)=\frac{\overline{W}_2(2)}{\overline{W}_2(6)}=\left(\frac{5\lg n}{\sqrt{\Delta\,\|p(x)\|^2}\,\sqrt{n}}\right)^{\frac{4}{5}},$$

$$R(2,7)=\frac{\overline{W}_2(2)}{\overline{W}_2(7)}=\left(\frac{\log_2 n+1}{\sqrt{\Delta\,\|p(x)\|^2}\,\sqrt{n}}\right)^{\frac{4}{5}}.$$

Если приведенные отношения меньше 1, то процедура дискретизации (2) более эффективна по сравнению с формулами (5) – (7). При этом их значения определяются коэффициентом $\sqrt{\Delta\,\|p(x)\|^2}$ процедуры дискретизации (2). Нетрудно убедиться, что данный коэффициент принимает постоянные значения для плот-

ностей вероятностей, вид которых не зависит от параметров закона распределения. К таким законам распределения относятся, например, равномерный, линейный, нормальный, экспоненциальный, Лапласа, для которых коэффициент $\sqrt{\Delta \|p(x)\|^2}$ принимает соответственно значения: 1; 1.15; 1.3; 1.7; 1.7. Данная закономерность не соблюдается для логнормального и трапециевидного законов распределения. Например, изменение параметров трапециевидного закона распределения меняет его вид от равномерного до треугольного распределений.

Исследовалась зависимость отношений $R(2,5)$, $R(2,6)$, $R(2,7)$ от объема n исходных данных для линейного $p_1(x)$, нормального $p_2(x)$ и экспоненциального $p_3(x)$ законов распределения. Отношение $R(2,5)<1$ и не зависит от n. Для плотностей вероятностей $p_1(x)$, $p_2(x)$, $p_3(x)$ его значения равны 0.89; 0.81; 0.65.

Анализ полученных отношений показывает, что при использовании метода дискретизации (2) достигаются более высокие аппроксимационные свойства регрессионной оценки плотности вероятности по сравнению с применением формул (5) – (7). Данный вывод согласуется с результатами исследований работы [7], в которой обосновывается оптимальный выбор количества интервалов дискретизации области определения плотности вероятности.

С ростом объема n исходных статистических данных преимущество метода дискретизации (2) над процедурами (6), (7) возрастает и особо проявляется при оценивании экспоненциального закона распределения, которому соответствует наибольшее из сравниваемых плотностей вероятностей значение коэффициента $\sqrt{\Delta \|p(x)\|^2}=1.7$.

Зависимости значений отношений $R(2, 6)$, $R(2, 7)$ от объема n исходных статистических данных для различных плотностей вероятностей представлены в таблице.

n	$p_1(x)$		$p_2(x)$		$p_3(x)$	
	R(2, 6)	R(2, 7)	R(2, 6)	R(2, 7)	R(2, 6)	R(2, 7)
40	1.08	0.89	0.98	0.81	0.79	0.65
60	0.99	0.82	0.91	0.74	0.73	0.59
80	0.94	0.76	0.85	0.69	0.69	0.56
100	0.89	0.72	0.81	0.65	0.65	0.53
120	0.86	0.69	0.78	0.62	0.63	0.50
140	0.83	0.66	0.75	0.60	0.60	0.48
160	0.80	0.64	0.73	0.58	0.59	0.47
180	0.78	0.62	0.71	0.56	0.57	0.45
200	0.76	0.60	0.69	0.55	0.55	0.44
220	0.74	0.59	0.67	0.53	0.54	0.43
240	0.72	0.57	0.66	0.52	0.53	0.42
260	0.71	0.56	0.64	0.51	0.52	0.41
280	0.70	0.55	0.63	0.50	0.51	0.40
300	0.68	0.54	0.62	0.49	0.50	0.39

Метод дискретизации Старджесса является менее эффективным в приведенных условиях исследований. Выбор количества интервалов дискретизации в соответствии с формулой Хайнкольда и Гаеде (5) более предпочтителен по сравнению с процедурой Брукса и Каррузера (6) при $n>100$.

Заключение

Методика синтеза регрессионной оценки плотности вероятности предполагает использование процедуры дискретизации области измерений случайной величины. Поэтому появляется возможность оценить эффективность различных формул дискретизации и на этой основе осуществить их анализ.

При восстановлении плотности вероятности случайной величины с линейным, нормальным, и экспоненциальным законами распределения целесообразно использовать метод дискретизации (2), который получен из анализа условий минимума среднеквадратической ошибки аппроксимации статистики (1). Менее предпочтительной при оценивании плотности вероятности является формула Старджесса. Применение формулы Хайнкольда и Гаеде при определении количества интервалов дискретизации области измерений случайной величины обладает преимуществом перед формулой Брукса и Каррузера при относительно больших значениях объема статистических данных $n > 100$.

Полученные результаты имеют также важное значение при доверительном оценивании плотности вероятности и проверки гипотез о распределениях случайных величин.

ЛИТЕРАТУРА

1. *Лапко А.В., Лапко В.А.* Непараметрические алгоритмы распознавания образов в задаче проверки статистической гипотезы о тождественности двух законов распределения случайных величин // Автометрия. – 2010. – № 6. – С. 47-53.
2. *Лапко А.В., Лапко В.А.* Сравнение эмпирической и теоретической функций распределения случайной величины на основе непараметрического классификатора // Автометрия. – 2012. – Т.48, № 1. – С. 45-49.
3. *Лапко А.В., Лапко В.А.* Непараметрические алгоритмы распознавания образов в задаче проверки гипотезы о распределениях случайных величин // Изв. вузов. Приборостроение. – 2011. – Т.54, № 4. – С. 67-72.
4. *Лапко А.В., Лапко В.А.* Непараметрические методики анализа множеств случайных величин // Автометрия. – 2003. – Т. 39, №1. – С.54-61.
5. *Лапко А.В., Лапко В.А.* Построение доверительных границ для плотности вероятности на основе ее регрессионной оценки // Метрология. – 2013. – №12. – С.3-9.
6. *Епанечников В.А.* Непараметрическая оценка многомерной плотности вероятности // Теория вероятности и ее применения. – 1969. – Т.14, №1. – С. 156-161.
7. *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.
8. *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, N 7. – C.763-767.
9. *Heinhold I., Gaede K.* Ingeniur statistic. – München: Wien, Springler Verlag, 1964.
10. *Шторм Р.* Теория вероятностей. Математическая статистика. Статистический контроль качества. – М.: Мир, 1970.
11. *Sturges H.A.* The choice of a class interval // J. American Statistical Association. – 1926. – P.65-66.

E-mail:
Лапко Александр Васильевич – lapko@icm.krasn.ru;
Лапко Василий Александрович – lapko@icm.krasn.ru.