

УДК 681.51:57

МОДЕЛЬ АГРЕГАЦИИ ОШИБОЧНЫХ ЗНАЧЕНИЙ КОЛИЧЕСТВЕННЫХ ПОКАЗАТЕЛЕЙ В МНОГОУРОВНЕВЫХ ИНФОРМАЦИОННЫХ СИСТЕМАХ

С.Я. Майстренко

Институт проблем математических машин и систем НАН Украины

e-mail: takam@post.com

Процесс решения многих задач в многоуровневых информационных системах связан со сбором первичных количественных показателей, их последовательной агрегацией (в частности, суммированием), накоплением и последующей обработкой. В связи с возможностью появления ошибок в первичных данных представляет интерес оценка ожидаемой точности агрегированных показателей. Под точностью здесь понимается математическое ожидание модуля смещения (отклонения в большую или меньшую сторону) расчетного значения агрегированного показателя, содержащего сумму некоторого количества ошибочных первичных значений, от его потенциально истинного значения, соответствующего сумме неискаженных данных.

Примем следующие обозначения и допущения:

M – количество составляющих слагаемых агрегированного показателя (агрегата);

m – количество ошибочных значений в агрегате, искаженных на случайную величину с математическим ожиданием $\pm u^{(1)}$ и дисперсией $d^{(1)}$ (единичное смещение и дисперсия);

n – количество разрядов суммируемых первичных показателей (будем считать его постоянным);

q – основание системы счисления в представлении показателя;

$U^{(m)}$ и $D^{(m)}$ – математическое ожидание и дисперсия модуля смещения суммы m искаженных показателей (смещение и дисперсия агрегата).

С учетом принятых обозначений расчетное значение агрегата $S(M)$ представляет собой случайную величину с математическим ожиданием $S(M) = S_0(M) \pm U^{(m)}$ и среднеквадратичным отклонением $\sigma^{(m)} = \sqrt{D^{(m)}}$, где $S_0(M)$ – истинное значение агрегата.

Оценка ожидаемого смещения значения агрегата

Схема агрегирования показателей, каждый из которых искажен на случайную величину с математическим ожиданием $\pm u^{(1)}$, показана на рис.1.

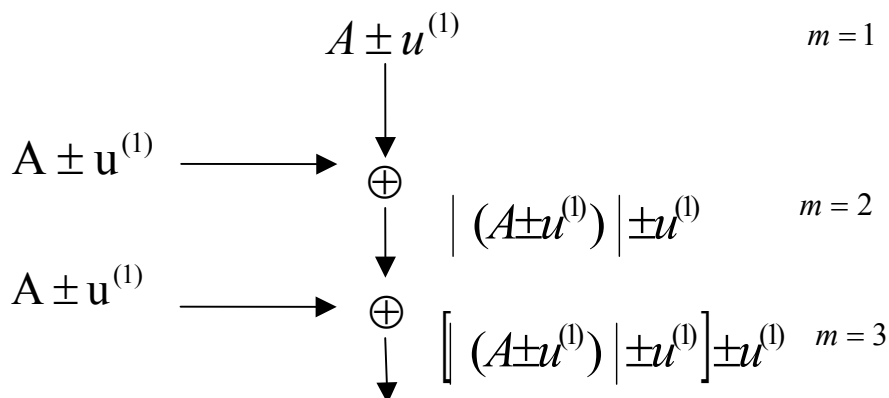


Рис.1

Обобщая приведенный пример на случай суммы m искаженных показателей, поставим в соответствие m показателям m -разрядное двоичное число. Припишем q -му разряду следующий смысл: если значение разряда равно 0, единичное смещение соответствующего показателя имеет знак "+", в противном случае единичное смещение имеет знак "-". В такой интерпретации, например для $m=8$ двоичному числу 01110111 соответствует случайное смещение

$$U_8 = | + 2u^{(1)} - 6u^{(1)} | = | - 4u^{(1)} | = 4u^{(1)}.$$

С учетом того, что общее количество всевозможных m - разрядных двоичных чисел равно 2^m , можно показать, что

$$U^{(m)} = k^{(m)} u^{(1)}, \quad (1)$$

где

$$k^{(m)} = \begin{cases} \frac{2}{m} \sum_{i=1}^{\frac{m}{2}} C_m^i (m-2i) & \text{для четных } m \\ \frac{2}{m} \sum_{i=1}^{\frac{m-1}{2}} C_m^i (m-2i) & \text{для нечетных } m \end{cases} \quad (2)$$

Соответственно

$$D^{(m)} = m d^{(1)} \quad (3)$$

В табл.1 приведены значения $k^{(m)}$, рассчитанные для $m = 1 \div 101$. Как видно из табл.1 с ростом m значение $U^{(m)}$ тоже растет, но значительно медленнее, чем m , - примерно в логарифмическом масштабе $\log_2 m$ (в очень грубом приближении). Так, например, увеличение m в 100 раз (с 1 до 100) вызывает увеличение $k^{(m)}$ только примерно в 8 раз.

Таблица 1

m	1	2	3	4	7	8	10	50	100	101
$k^{(m)}$	1.00	1.00	1.50	1.50	2.19	2.19	2.46	5.61	7.96	8.04

Единичное смещение

Определяя значение $u^{(1)}$, ограничимся предположением, что ошибочное значение n -разрядного показателя возникло в результате искажения истинного значения однократной ошибкой, как наиболее вероятной. Предполагая, что в n -разрядном показателе может быть искажен любой из n разрядов с вероятностью $\frac{1}{n}$, и учитывая, что вклад разряда с позицией j ($j = 1, \dots, n$) в значение n -разрядного числа пропорционален q^{j-1} , в результате простых рассуждений получим:

$$u^{(1)} = \frac{2 \sum_{i=1}^{q-1} \sum_{j=1}^n i(q-i) q^{j-1}}{nq(q-1)}. \quad (4)$$

Соответственно

$$d^{(1)} = \frac{2 \sum_{i=1}^{q-1} \sum_{j=1}^n [(q-i) q^{j-1} - u^{(1)}]^2 \cdot i}{nq(q-1)} \quad (5)$$

В табл.2 приведены расчетные данные для $u^{(1)}$, $d^{(1)}$ и среднеквадратичного отклонения $\sigma^{(1)} = \sqrt{d^{(1)}}$ для $q = 10$.

Таблица 2

n	1	4	6	8	10	12
$u^{(1)}$	3.67	$1,02 \cdot 10^3$	$6.79 \cdot 10^4$	$5.09 \cdot 10^6$	$4.07 \cdot 10^8$	$3.39 \cdot 10^{10}$
$d^{(1)}$	4.89	$3.59 \cdot 10^6$	$2.62 \cdot 10^{10}$	$2.05 \cdot 10^{14}$	$1.68 \cdot 10^{18}$	$1.42 \cdot 10^{22}$
$\sigma^{(1)}$	2.21	$1.89 \cdot 10^3$	$1.62 \cdot 10^5$	$1.43 \cdot 10^7$	$1.29 \cdot 10^9$	$1.19 \cdot 10^{11}$

Можно показать, что с повышением q значение $u^{(1)}$ асимптотически приближается к $0,33 \dots q$. Это вытекает из следующей “непрерывной” модели. Представим себе непрерывную числовую ось $0 \div Q$. По этой оси случайным образом от координаты x к координате y перемещается некий “курсор”. Длина перемещения l_{xy} курсора является случайным числом в интервале $0 \div Q$. Математическое ожидание $\bar{l}_{xy}$ длины перемещения, как легко показать, равно:

$$\bar{l}_{xy} = u_Q^{(1)} = \frac{1}{Q^2} \int_0^Q \int_0^Q |x - y| dx dy = 0,33 \dots Q.$$

Поскольку значение Q в этой модели можно трактовать, как q^n , значение $0,33 \dots q$ могло бы быть использовано для уточненной оценки взвешенного значения $u^{(1)}$ (с дополнительным учетом удельного веса многократных ошибок). Тем не менее, для получения ориентировочных оценок мы ограничимся упрощенной моделью, основанной на первоначальном предположении об искажении показателей однократными ошибками.

Анализ полученных зависимостей

Для адекватной и наглядной интерпретации выражений (1) – (5) и иллюстрирующих их данных табл. 1,2 рассмотрим оценки относительных значений $U^{(m)}$ и $\sigma^{(m)}$ по отношению к среднему значению агрегата $\bar{S}(M)$.

Положим $\bar{S}(M) = 0,5Mq^n$, $m = \gamma M$ и $q = 10$.

Тогда

$$U_{отн}^{(m)} = \frac{U^{(m)}}{0,5 \cdot M \cdot q^n} = \frac{k^{(\gamma M)} \cdot u^{(1)}}{0,5 \cdot M \cdot 10^n} \quad (6)$$

$$\sigma_{отн}^{(m)} = \frac{\sqrt{D^{(\gamma M)}}}{0,5 \cdot M \cdot q^n} = \frac{\sqrt{\gamma M d^{(1)}}}{0,5 \cdot M \cdot 10^n} \quad (7)$$

Таблица 3

M m		100 2	200 3	400 6	600 9	800 12	1000 15
$n = 3$	$U_{отн}^{(m)}$		2,0350E-3	1,2719E-3	1,1129E-3	9,1813E-4	8,5255E-4
	$D_{отн}^{(m)}$		1,2997E+0	1,2997E+0	1,2997E+0	1,2997E+0	1,2997E+0
	$\sigma_{отн}^{(m)}$		3,6051E-3	2,5492E-3	2,0814E-3	1,8026E-3	1,6123E-3
$n = 4$	$U_{отн}^{(m)}$	2,0368E-3	1,5276E-3	1,1139E-3	9,1896E-4	7,9999E-4	7,1777E-4
	$D_{отн}^{(m)}$	1,4370E+1	1,4370E+1	1,4370E+1	1,4370E+1	1,4370E+1	1,4370E+1
	$\sigma_{отн}^{(m)}$	5,3609E-3	3,7908E-3	2,6805E-3	2,1886E-3	1,8954E-3	1,6953E-3
$n = 6$	$U_{отн}^{(m)}$	2,0370E-3	1,2731E-3	9,1905E-4	7,5562E-4	6,5666E-4	5,8856E-4
	$D_{отн}^{(m)}$	1,5752E+3	1,5752E+3	1,5752E+3	1,5752E+3	1,5752E+3	1,5752E+3
	$\sigma_{отн}^{(m)}$	5,6129E-3	3,9689E-3	2,8064E-3	2,2914E-3	1,9845E-3	1,7749E-3

В табл.3 приведены расчетные значения $U_{отн.}^{(m)}$, $D_{отн.}^{(m)}$, $\sigma_{отн.}^{(m)}$ для значений $q = 10$; $n = 3, 4, 6$; $\frac{\gamma}{n} = 5 \cdot 10^{-3}$.

1) Ожидаемое относительное смещение значения агрегата медленно убывает с ростом n и M , оставаясь при этом весьма небольшим. Так для $n = 6$, $M = 100 - 1000$, $m = 3 - 30$ значение $U_{отн.}^{(m)}$ составляет всего $0,2\% \div 0,059\%$.

2) Относительное среднеквадратичное отклонение также убывает с ростом M, n (пропорционально $\sqrt{m}$) и, как и следовало бы ожидать, сравнительно велико. Оценки предельных величин фактических значений смещения агрегата относительно математического ожидания могут быть получены на основе известного неравенства Чебышева [3]:

$$p(t) \leq \frac{\sigma^2}{t^2}, \quad (8)$$

где $p(t)$ - вероятность отклонения значения случайной величины (в данном случае, смещения) от математического ожидания более, чем на t единиц;
 σ - среднеквадратическое отклонение.

Таким образом, полученные соотношения и иллюстрирующие их данные позволяют ориентировочно оценить ожидаемые значения отклонения агрегата от истинного значения с учетом конкретных условий, связанных в частности, с технологией формирования первичных данных, контроля достоверности и дальнейшей агрегации. Такая оценка может быть полезна в системах поддержки, принятия решений, ориентированных на различные проблемные области (вооруженные силы, экономика, социология и др.). Полученные оценки могут быть уточнены путем учета удельного веса многократных ошибок при определении $u^{(1)}$.

Основной вклад в математическое ожидание и дисперсию смещения дает, естественно, самый старший разряд показателя. Поэтому специальная дополнительная защита старшего разряда, повышающая вероятность обнаружения его искажения, позволила бы уменьшить как $U^{(m)}$, так и $D^{(m)}$. Такая защита могла бы быть реализована, в частности, повторением значения этого разряда при формировании данных или вводе. При этом, например, число 5734 записывалось бы, как 55734 или может быть 57345. Исследование такого случая могло бы составить отдельную задачу. Однако практическая его реализация возможна, по-видимому, только в каких-то специальных случаях.

ЛИТЕРАТУРА

1. Пивоваров А.Н. Методы обеспечения достоверности информации в АСУ: (Обзор методов и фактические данные). — М. : Радио и связь, 1982. — 144 с.
2. Литвинов В.А., Крамаренко В.В. Контроль достоверности и восстановление информации в человеко-машинных системах. Киев. Техника, 1988г., 200с.
3. В. Феллер Введение в теорию вероятностей и ее приложения. Москва. «Мир 1967г. 498с.
4. Горбань І.І. Теорія ймовірностей і математична статистика //Київ: Інститут проблем математичних машин та систем, 2003.- 244 с.