УДК 519.23+519.25

Выявление выбросов в методе максимума согласования при анализе интервальных данных

С.П. Шарый
ИВТ СО РАН, г. Новосибирск

Наша работа посвящена анализу данных, заданных неточно и имеющих интервальную неопределённость. Рассматривается задача восстановления зависимостей по интервальным данным, которая в последние десятилетия привлекала стабильное внимание специалистов. Для её решения был предложен ряд методов, в частности, так называемый метод максимума согласования [1–5], основанный на максимизации специальных функционалов, которые дают количественную меру согласования (совместности) параметров зависимости и интервальных данных измерений.

Далее в работе мы рассматриваем простейшую зависимость вида

$$b = x_1 a_1 + x_2 a_2 + \ldots + x_n a_n , \qquad (1)$$

в которой значения b являются линейной функцией от независимых переменных $a_1, a_2, \ldots, a_n$. Необходимо определить неизвестные коэффициенты $x_1, x_2, \ldots, x_n$, чтобы получившаяся функциональная зависимость «наилучшим образом» соответствовала заданному набору значений $a_1, a_2, \ldots, a_n$ и b, полученному в результате m измерений (наблюдений). При этом результаты измерений не известны точно, и нам даны лишь интервалы их возможных значений. Таким образом, результат i-го измерения j-ой независимой переменной $a_j^{(i)}$ принадлежит интервалу $\boldsymbol{a}_j^{(i)}$, а результат i-го измерения $b^{(i)}$ зависимой переменной принадлежит интервалу $\boldsymbol{b}^{(i)}$, где $i = 1, 2, \ldots, m$, $j = 1, 2, \ldots, n$. В целом имеем набор интервальных данных

$$\begin{array}{c} \boldsymbol{a}_1^{(1)}, \boldsymbol{a}_2^{(1)}, \ldots, \boldsymbol{a}_n^{(1)}, \boldsymbol{b}^{(1)}, \\ \boldsymbol{a}_1^{(2)}, \boldsymbol{a}_2^{(2)}, \ldots, \boldsymbol{a}_n^{(2)}, \boldsymbol{b}^{(2)}, \\ \ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots \\ \boldsymbol{a}_1^{(m)}, \boldsymbol{a}_2^{(m)}, \ldots, \boldsymbol{a}_n^{(m)}, \boldsymbol{b}^{(m)}, \end{array} \qquad (2)$$

по которым строится интервальная система линейных уравнений для нахождения параметров зависимости.

В методе максимума согласования оценкой параметров восстанавливаемой линейной зависимости берётся вектор $(x_1, x_2, \ldots, x_n)^\top$, на котором достигается максимум так называемых распознающих функционалов тех или иных множеств решений интервальной системы уравнений, построенной по данным измерений. Это функции, которые задаются либо выражением

$$\mathrm{Uss}\,(x, \boldsymbol{A}, \boldsymbol{b}) = \min_{1\le i\le m}\Big\{ \operatorname{rad} \boldsymbol{b}_i + \sum_{1\le j\le n} (\operatorname{rad} \boldsymbol{a}_{ij})\,|x_j| - \big|\operatorname{mid} \boldsymbol{b}_i - \sum_{1\le j\le n} (\operatorname{mid} \boldsymbol{a}_{ij})\,x_j\big|\Big\} \tag{3}$$

в случае обычного (слабого) согласования параметров и данных (см. [2, 3, 4]), либо выражением

$$\mathrm{Tol}\,(x, \boldsymbol{A}, \boldsymbol{b}) = \min_{1\le i\le m}\Big\{ \operatorname{rad} \boldsymbol{b}_i - \big|\operatorname{mid} \boldsymbol{b}_i - \sum_{1\le j\le n} \boldsymbol{a}_{ij}\,x_j\big|\Big\} \tag{4}$$

в случае сильного согласования параметров и данных (см. [5]).

При восстановлении зависимостей важной практической задачей является выявление *выбросов* – таких измерений, результат которых выделяется из общей выборки и никак не характеризует искомую функциональную зависимость. Цель этой работы – изложение простого полуэвристического приёма для выявления измерений, подозрительных на выбросы, в рамках общей вычислительной схемы метода максимума согласования.

Исходный пунктом нашей методики является то простое наблюдение, что выражения для распознающих функционалов имеют весьма специальный вид, в котором окончательное значение получается как минимум от значений ряда выражений одинаковой структуры (стоящих внутри фигурных скобок в (3)–(4)), которые вычисляются по строкам матрицы данных (2). Мы будем называть их *образующими* распознающих функционалов. Фактически, их значения в точке $(x_1, x_2, \ldots, x_n)^\top$ характеризуют отдельные измерения, давая для каждого из них меру согласования (совместности) данных в этом измерении с вектором параметров $(x_1, x_2, \ldots, x_n)^\top$.

С другой стороны, выбросы – это измерения, удаление которых резко увеличивает меру согласования оставшейся части выборки.

Как следствие, приходим к следующей естественной идее. В точке максимума распознающего функционала нужно посмотреть на значения его образующих, соответствующих отдельным измерениям, и если какие-то из этих образующих существенно меньше остальных, то они и являются кандидатами на выбросы.

Высказанная идея верна по сути, но на пути её успешного применения стоят некоторые принципиальные ограничения, которые следует учитывать при интерпретации результатов расчётов.

Напомним, что в пределе, когда интервалы неопределённости данных вырождаются в точки и мы должны восстанавливать зависимость по точным данным, метод максимума согласования (как слабая, так и сильная версии) переходит в чебышёвское сглаживание данных [4, 5], т.е. в их приближение в равномерной метрике. Один из основных результатов теории равномерного приближения функций — это знаменитая

Теорема Чебышёва [6, 7]. Для того, чтобы многочлен n-ой степени $P(x)$ являлся многочленом наилучшего равномерного приближения непрерывной на интервале $[a, b]$ функции $f(x)$, необходимо и достаточно, чтобы на $[a, b]$ существовали по крайней мере $(n+2)$ точки $x_0 < x_1 < \ldots x_n < x_{n+1}$, такие что разность $f(x_i) - P(x_i)$, $i = 0, 1, \ldots, n+1$, принимает в них равные по абсолютной величине значения, которые последовательно меняют знак от точки к точке.

Точки $x_0 < x_1 < \ldots < x_n < x_{n+1}$, о которых идёт речь в теореме Чебышёва, называются, как известно, точками *чебышёвского альтернанса*. Если ищется наилучшее равномерное приближение линейной функцией, т. е. полиномом первой степени $n = 1$, то $n+2 = 3$, так что точек альтернанса должно быть не менее трёх штук. Но нередко их бывает гораздо больше. Нетрудно понять, что точки альтернанса соответствуют тем измерениям, значения образующих для которых – наименьшие, и из сделанного наблюдения следует, что таких точек не может одна или две. Их принципиально не меньше трёх, а, вообще говоря, и больше.

Что происходит в случае интервальных данных? Вместо точек мы имеем брусы неопределённости измерений в пространстве $\mathbb{R}^{n+1}$, так что в общем случае теорема Чебышёва здесь, строго говоря, неприменима. Тем не менее, если интервалы данных «не слишком широки» (или «достаточно узки»), то теорема Чебышёва всё-таки остаётся верной, и мы можем считать, что количество точек альтернанса остаётся равным как минимум $n+2$, т.е. 3 в линейном случае. Опять-таки, в реальных ситуациях их может быть довольно много, что хорошо демонстрируется при работе с практическими задачами.

Таким образом, в методе максимума согласования выбросы, если они имеются, в силу принципиальных математических причин всегда маскируются обычными информативными измерениями.

Тем не менее, если количество обрабатываемых измерений велико, то любая дополнительная информация о выбросах, любая техника, позволяющая сузить «круг подозреваемых», может оказаться полезной и имеет смысл быть применённой. Особенно, когда затраты на её реализацию пренебрежимо малы, как это имеет место с предложенной выше методикой исследования образующих распознающего функционала в точке максимума.

Библиографический список

1. Шарый С.П. Разрешимость интервальных линейных уравнений и анализ данных с неопределённостями // Автоматика и Телемеханика. – 2012. – №2 – С. 111–125.

2. Шарый С.П., Шарая И.А. Распознавание разрешимости интервальных уравнений и его приложения к анализу данных // Вычислительные Технологии. – 2013. – Т. 18, №3. – С. 80–109.

3. Kreinovich V., Shary S.P. Interval methods for data fitting under uncertainty: a probabilistic treatment // Reliable Computing. – 2016. – Vol. 23. – P. 105–140.

4. Шарый С.П. Метод максимума согласования для восстановления зависимостей по данным с интервальной неопределённостью // Известия Академии Наук. Теория и системы управления. – 2017. – №6. – С. 3–19.

5. Шарый С.П. Сильная согласованность в задаче восстановления зависимостей при интервальной неопределённости данных // Вычислительные Технологии. – 2017. – Т. 22, №2. – С. 150–172.

6. Бахвалов Н.С., Жидков Н.П., Кобельков Г.Н. Численные методы. – Москва: Бином-Лаборатория базовых знаний, 2003.

7. Натансон И.П. Конструктивная теория функций. – Москва-Ленинград: ГИТТЛ, 1949.