УДК 004.021; 519.2

Особенности реализации алгоритма многомерного взвешивания наблюдений

В. Л. Чечулин, А. С. Кичёв
Пермский государственный национальный исследовательский университет
Россия, 614990, г. Пермь, ул. Букирева, 15
chechulinvl@mail.ru; 8 (342) 2-396-424

Описана реализация алгоритма многомерного взвешивания наблюдений, использующая исходную функцию взвешивания, основанную на неравенстве Чебышёва, и итерационный подход, – результаты взвешивания наблюдений предназначены для получения устойчивых оценок в многомерных методах (корреляции, регрессии и т. п.), приведены примеры вычисления таких оценок. Приведенный алгоритм используется при итерациях перенормировки по одномерным подпространствам, т. е. не зависим от выбора масштаба измерений по отдельным осям исходных многомерных данных.

Ключевые слова: *алгоритм взвешивания наблюдений; неравенство Чебышёва; перенормировки по осям; итерационное взвешивание; устойчивая корреляция; устойчивая регрессия.*

DOI: 10.17072/1993-0550-2017-4-75-78

Предисловие

Необходимость построения статистических оценок, позволяющих фильтровать шумы, была отмечена еще в последней трети XX в. [2]. Подходы, использовавшиеся в то время, ограничивались предположением, что исходные (незашумленные) данные имеют известное (обычно нормальное распределение), для которого строилась функция влияния, позволяющая отсекать выбросы наблюдений [3], [4].

Однако вид функции распределения не всегда известен, и тем более не всегда можно предполагать нормальность распределения незашумленных данных, поэтому с начала XXI в. для оценок плотности вероятности (взвешивания наблюдений стали применять более обобщенный тип функций, нежели предполагаемые для нормального распределения); так, в [1] отмечается, что используется "некоторая четная положительная функция, удовлетворяющая следующим условиям:

$$\int_{-\infty}^{\infty} K(x)dx = 1, \quad \int_{-\infty}^{\infty} K^2(x)dx < \infty$$" [1, с. 119].

Дальнейшие исследования оснований устойчивого оценивания показали, что в качестве взвешивающей функции, не зависящей от типа распределения оцениваемой совокупности, фундаментально обосновано применение функции, использующей неравенство Чебышёва [6].

Для построения одномерных оценок такой способ взвешивания был успешно проверен (см. список литературы в [6]).

Для многомерных оценок (корреляции, регрессии и т. п.) возникала проблема разномасштабности отдельных измерений (одномерных подпространств) исходных данных. Проблема эта решалась в частном случае для сопоставимых масштабов данных, см. [7].

Ниже приведен универсальный вариант алгоритма многомерного взвешивания, использующий перенормировки и итерационный процесс для независимости масштабов данных по осям (на примере двумерного случая).

1. Многомерный (двумерный) метод взвешиваняи

Имеется двумерный массив наблюдений A_i, где i= 1, n, координаты наблюдений пары (x_i, y_i).

На первом шаге алгоритма вычисляется среднее $\bar{x}$ $\bar{y}$ и стандартное отклонение σ(x) σ(y) для каждой из переменных, а затем выполняем перенормировку для каждой из координат:

$$x^*_i = (x_i - \bar{x}) / \sigma(x),$$
$$y^*_i = (y_i - \bar{y}) / \sigma(y).$$

На втором шаге вычисляются посредством функции взвешивания и первые приближения весов. Для чего точки в неренормированных координатах (x*, y*) рассматриваются попарно в цикле по i и по j, для каждой пары точек вычисляется евклидово расстояние между точками $r_{ij}(A_i, A_j)$, которое подается входным параметром в функцию взвешивания f(i, j, σ):

$$r_{ij} = \sqrt[2]{(x^*_i - x^*_j)^2 + (y^*_i - y^*_j)^2} \quad .$$

Поскольку проведена перенормировка, то σ(x*)=σ(y*)=1 и функция взвешивания имеет вид

$$f(i, j) = \begin{cases} 1/r^2_{ij}, \text{ если } r_{ij} > 1 \\ r_{ij}, \text{ если } r_{ij} \le 1 \end{cases} .$$

В этом же цикле вычисляются первые приближения весов (для каждого i при переменном j):

$$\omega_i := 0,$$
$$\omega_i := \omega_i + f(i, j).$$

Затем веса перенормируются:

$$\omega^*_i = \omega_i / \Sigma_{i=1,n}\, \omega_i ,$$

и по этим весам вычисляются устойчивые средние и устойчивые отклонения для каждой координаты (схема обработки данных приведена на рис. 1).

Основным отличием данной реализации алгоритма является то, что исходные данные являются глобальными переменными, а веса и перенормировки данных посредством взвешенных оценок – локальными переменными. То есть, оценивание выполняется по исходным данным, а перенормировки данных носят лишь служебный вид, используются для вычисления весов наблюдений на очередном шаге итерационного процесса.

Рис. 1. *Схема обработки данных (исходные данные – глобальные переменные, веса – локальные переменные)*

Процесс взвешивания возвращается к первому шагу с тем лишь отличием, что перенормировка выполняется с использованием устойчивых характеристик (устойчивое среднее и отклонение):

$$x^*_i = \frac{x_i - \overline{x_{устойчив.}}}{\sigma_{устойчив.}(x)} .$$

И для перенормированных значений находятся новые веса. Оценки же выполняются для исходных значений переменных с использованием найденных весов.

$$\overline{x_{уст.}} = \sum_{i=1}^{n} x_i \cdot \omega_i ,$$

$$\sigma_{уст.} = \sqrt{\sum_{i=1}^{n} (x_i - \overline{x_{уст.}})^2 \cdot \omega_i} ,$$

$$\sigma_{уст.} = \sqrt{\frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} (x_i - x_j)^2 \cdot \omega_i \cdot \omega_j} . \quad (1)$$

Таким образом, выполняется итерационное уточнение полученных весов наблюдений. Итерационный процесс продолжается до тех пор, пока норма вектора – разницы весов на текущем и предыдущем шаге не станет меньше заданной величины.

После того как реализован алгоритм взвешивания для многомерного случая, становится возможным рассчитывать устойчивый коэффициент корреляции и построить устойчивые регрессионные прямые (двумерный случай).

2. Устойчивый коэффициент корреляции

Взвешенная оценка ковариации находится по формуле [5]:

$$\mathrm{cov}_w(X,Y) = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}(x_i - x_j)\cdot(y_i - y_j)\cdot\omega_i\cdot\omega_j.$$

Отсюда получаем взвешенный коэффициент корреляции:

$$corr_w(X,Y) = \frac{\mathrm{cov}_w(X,Y)}{\sqrt{D_w(X)\cdot D_w(Y)}},$$

где D_w – взвешенная оценка дисперсии, которая находится по формуле следующей [5], см. также формулу (1):

$$D_\omega(X) = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}(x_i - x_j)^2\cdot\omega_i\cdot\omega_j\ .$$

3. Устойчивая регрессия

Уравнение регрессии в стандартном виде выглядит следующим образом:

$$\hat{y}_i = a + b\cdot x_i.$$

Выведем коэффициенты a и b для устойчивой регрессии. При помощи МНК минимизируется величина, в которую добавлены веса ω_i:

$$\sum_{i=1}^{n}((y_i - \hat{y}_i)^2\cdot\omega_i) \to \min.$$

Подставим уравнение регрессии:

$$\sum_{i=1}^{n}((y_i - a - bx_i)^2\,\omega_i) \to \min.$$

Полученная функция минимизируется:

$$Q(a,b) = \sum_{i=1}^{n}((y_i - a - bx_i)^2\,\omega_i).$$

Необходимое условие минимума:

$$\begin{cases}\dfrac{\partial Q}{\partial a} = 0;\\[2ex] \dfrac{\partial Q}{\partial b} = 0.\end{cases}$$

Получаем систему:

$$\begin{cases}2\cdot\sum_{i=1}^{n}(y_i - a - bx_i)(-\omega_i x_i) = 0;\\ 2\cdot\sum_{i=1}^{n}(y_i - a - bx_i)(-\omega_i) = 0.\end{cases}$$

Система преобразуется в

$$\begin{cases}\sum_{i=1}^{n} y_i\omega_i x_i = b\cdot\sum_{i=1}^{n}\omega_i\,x_i^2 + a\cdot\sum_{i=1}^{n}\omega_i\,x_i;\\ \sum_{i=1}^{n} y_i\omega_i = b\cdot\sum_{i=1}^{n}\omega_i\,x_i + a\cdot\sum_{i=1}^{n}\omega_i\,.\end{cases}$$

Разрешив систему, получим

$$a = \sum_{i=1}^{n} y_i\omega_i - b\cdot\sum_{i=1}^{n} x_i\omega_i$$

$$b = \frac{\sum_{i=1}^{n} y_i x_i\omega_i - \sum_{i=1}^{n} y_i\omega_i\cdot\sum_{i=1}^{n} x_i\omega_i}{\sum_{i=1}^{n} x_i^2\omega_i - (\sum_{i=1}^{n} x_i\omega_i)^2}.$$

Или в другой записи:

$$a = (\overline{Y})_\omega - b\cdot(\overline{X})_\omega,\ b = \frac{(\overline{XY})_\omega - (\overline{Y})_\omega(\overline{X})_\omega}{(\overline{X^2})_\omega - (\overline{X})_\omega^2},$$

где $(\overline{X})_\omega, (\overline{Y})_\omega, (\overline{XY})_\omega, (\overline{X^2})_\omega$ – устойчивые средние величин *X, Y, XY, X*^{2}.

4. Сравнение оценок со стандартными

Для сравнения возьмем искусственный пример данных (см. таблицу).

Пример данных для тестирования методов

№	X	Y
1	1	1
2	2	2
3	3	3
4	4	4
5	5	5
6	6	6
7	7	7
8	8	8
9	9	9
10	10	10
11	11	11
12	12	12
13	13	13
14	14	14
15	15	15
16	-1	100

Заметим, что точки 1–15 линейно зависимы (коэффициент корреляции 1), однако с удаленной точкой № 16 коэффициент корреляции Пирсона = -0,29. Взвешенный коэффициент корреляции для этих же данных = 0,88. Построены также линии регрессии (см. рис. 2).

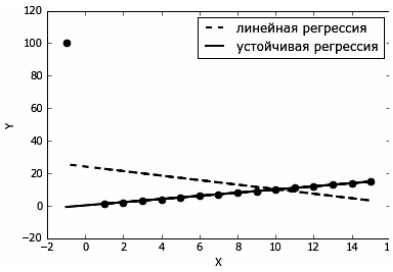

Рис. 2. *Сравнение стандартной линейной регрессии с устойчивой*

Из приведенного примера (рис. 2) видно, что программно реализованные устойчивые методы "улавливают" основную закономерность, несмотря на сильно отклоняющуюся точку (выброс), тогда как стандартные методы для данной задачи слабо подходят.

Заключение

Таким образом, приведенный алгоритм взвешивания с двумерного случая прозрачно обобщаем на многомерный, для таких статистических методов, как метод главных компонент, корреляции многомерных данных, парные регрессии многомерных данных и т. п.

Многочисленные конкретные приложения разработанного алгоритма к анализу данных в различных предметных областях подлежат отдельному описанию.

Список литературы

1. *Алексеев В.Г.* О допустимых непараметрических оценках плотности вероятности // Автометрия. 2005. Т. 41, № 3. С. 118–121.
2. *Мостеллер Ф., Тьюки Дж.* Анализ данных и регрессия: в 2 вып. / пер. с англ. Ю.Н. Благовещенского; под ред. и с предисл. Ю.П. Адлера. М.: Финансы и статистика, 1982. Вып. 1. 317 с.
3. *Хампель Ф., Рончетти Э., Рауссеу П. и др.* Робастность в статистике. Подход на основе функций влияния / пер. с англ. В.М. Золотарева. М.: Мир, 1989. 512 с.
4. *Хьюбер Дж. П.* Робастность в статистике / пер. с англ. И.А. Маховой, В.И. Хохлова; под ред. И.Г. Журбенко. М.: Мир, 1984. 304 с.
5. *Чечулин В.Л.* О взвешенной оценке масштаба (дисперсии) выборки, не использующей оценку положения (среднего) // *Чечулин В.Л.* Ст. в журнале "Университетские исследования" 2009–2014 гг.: [Электронный ресурс] сб. / Перм. гос. нац. исслед. ун-т. Электрон. дан. Пермь, 2015. С. 244–246.
6. *Чечулин В.Л.* Обоснование взвешивания наблюдений посредством неравенства Чебышёва // Чечулин В.Л. Статьи разных лет. 2016. Вып. 3 / Перм. гос. нац. исслед. ун-т. Пермь, 2016. С. 27–32.
7. *Чечулин В.Л., Грацилёв В.И.* Устойчивое регрессионное оценивание, основанное на неравенстве Чебышёва // Чечулин В.Л. Ст. в журнале "Университетские исследования" 2009–2014 гг.: [Электронный ресурс] сб. / Перм. гос. нац. исслед. ун-т. Электрон. дан. Пермь, 2015.

Multidimensional weighting algorithm: features of the implementation

V. L. Chechulin, A. S. Kichev
Perm State University; 15, Bukireva st., Perm, 614990, Russia
chechulinvl@mail.ru; 8 (342) 2-396-424

The paper describes implementation of an algorithm for multidimensional weighting of observations with the use of the original weighting function based on the Chebyshev inequality and the iterative approach. The results of weighing the observations are intended to obtain stable estimates in multidimensional methods (correlations, regressions, etc.); the examples of such estimates calculation are given. During iterations, the above algorithm uses renormalization over one-dimensional subspaces, i.e., it is independent of the measurement scale chosen for the individual axes of the initial multidimensional data.

Keywords: *weighting algorithm; Chebyshev's inequality; renormalization along the axes; iterative weighing; robust correlation; robust regression.*