Раздел II. Информатика

И.М. Никольский, К.К. Фурманов

ПАРАЛЛЕЛЬНЫЙ АЛГОРИТМ ПОИСКА СТРУКТУРНЫХ ИЗМЕНЕНИЙ ВО ВРЕМЕННЫХ РЯДАХ

1. Введение.

Под структурным изменением понимается изменение параметров модели временного ряда. Поиск точек, где происходят такие изменения, – важная задача во многих областях науки и техники. Так, например, проблема позиционирования (place recognition) в компьютерном зрении может быть решена с помощью детектирования структурных изменений во входящем видеопотоке[1]. В работе [2] по точкам сдвига среднего изучается взаимосвязь между биржевыми индексами США, Японии и Европы. В генетике большой интерес вызывает поиск изменений композиционной структуры генома [3].

В данной работе рассматривается особый вид структурных изменений – сдвиг среднего, т.е. изменение математического ожидания при неизменной дисперсии.

На данной момент хорошо развита теория поиска одной точки сдвига временного ряда. Начало было положено Шухартом в статье 1925го года [4]. Большой вклад внесли Пейдж[5], Ширяев[6], Робертс[7], Никифоров [8] и другие исследователи. Были разработаны методы, основанные на контрольных картах (CUSUM, EWMA, карты Шухарта), информационных критериях, байесовском подходе, максимальном правдоподобии.

Решению задачи поиска всех точек сдвига посвящено значительно меньше работ. Здесь наиболее широко используемым является, по всей видимости, метод бинарной сегментации (binary segmentation) [9]. Конкуренцию ему составляют методы, использующие динамическое программирование (метод PELT[10], оптимального разбиения [11] и др.).

Задачи обработки больших данных (big data), возникающие в таких областях как генетика, финансы, дают новый импульс для поиска новых методов детектирования сдвига среднего. Обработку временных рядов большой длины невозможно проводить без использования многопроцессорных ЭВМ (суперкомпьютеров), что приводит к возникновению нового требования к соответствующим алгоритмам - они должны обладать высокой степенью параллелизма.

В данной работе предложен новый параллельный алгоритм поиска точек сдвига среднего временного ряда. Основной его идеей является разбиение временного ряда на сегменты одинаковой длины. Поиск точки сдвига на каждом сегменте проводится независимо в соответствии с моделью данных, описанной в пункте 2. В пункте 3 приводится статистика с помощью которой производится проверка статистической гипотезы о наличии точки сдвига. Плохая детектируемость точек у границ сегмента приводит к необходимости исследования окрестностей границ между сегментами сегментов (см. п. 4). Параллельная реализация алгоритма для машин с общей памятью описана в пункте 5.

2. Модель точки сдвига среднего.

Пусть дан временной ряд $\{x_i\}_{i=1}^{n} = x_1,..,x_n$. Пусть элемент выборки x_i $(1 \le i \le n)$ - реализация сл. вел. X_i. Предположим, что все X_i распределены нормально с одинаковой дисперсией σ^2. Требуется проверить гипотезу H_0:

$$E(X_1) = E(X_2) = ... = E(X_n).$$

Отвержение этой гипотезы будет означать, что существует такое k $(1 < k < n)$, что

$$X_i \propto N(\mu_1, \sigma^2), \ i = 1,..,k,$$
$$X_i \propto N(\mu_2, \sigma^2), \ i = k+1,..,n.$$

Здесь k - момент сдвига среднего, оценку которого требуется найти, если H_0 отвергнута.

3. Поиск одной точки сдвига среднего.

Наш метод поиска сдвигов среднего основан на разбиении временного ряда на сегменты одинаковой длины и поиска одного сдвига на каждом из сегментов. Ниже описывается процедура поиска точки сдвига на сегменте определенной длины *n*.

3.1 Тест для проверки гипотезы о наличии точки сдвига.

Для проверки гипотезы (см. п.2) будем использовать стандартную t-статистику, применяемую для проверки равенства средних двух выборок различного размера с одинаковой дисперсией. Здесь мы следуем [12]. Для каждого возможного положения точки сдвига k $(1 < k < n)$ вычислим следующим величины:

$$\tilde{\mu}_{kn}^{(1)} = \frac{\sum_{i=1}^{k} x_i}{k}, \ \tilde{\mu}_{kn}^{(2)} = \frac{\sum_{i=k+1}^{n} x_i}{n-k} \text{ - средние до и после сдвига,}$$

$$V_{kn} = \sum_{i=1}^{k} \left(x_i - \tilde{\mu}_{kn}^{(1)} \right)^2 + \sum_{i=k+1}^{n} \left(x_i - \tilde{\mu}_{kn}^{(2)} \right)^2 \text{ - сумма квадратов отклонений,}$$

$$\tilde{\sigma}_{kn}^2 = V_{kn} / (n-2) \text{ - выборочная оценка дисперсии.}$$

Стандартная t-статистика двухвыборочного критерия Стьюдента для выборок $\{x_i\}_{i=1}^{k}$ и $\{x_i\}_{i=k+1}^{n}$ имеет следующий вид:

$$T_{kn} = \sqrt{\frac{k(n-k)}{n}} \frac{\tilde{\mu}_{kn}^{(1)} - \tilde{\mu}_{kn}^{(2)}}{\tilde{\sigma}_{kn}}$$

Сам по себе критерий Стьюдента используется в том случае, когда сравниваемые выборки уже сформированы: известно, какие наблюдения входят в каждую из них, неизвестно лишь, совпадают ли в выборках математические ожидания. Мы, однако, не знаем точку сдвига, разделяющую куски ряда с различными математическими ожиданиями, и подбираем её так, чтобы статистика Стьюдента была наибольшей, что свидетельствовало бы о статистически наиболее достоверном различии. Получаем следующую статистику для проверки H_0(впервые предложена в [13]):

$$T_{\max,n} = \max_{1 \le k \le n-1} |T_{kn}|.$$

Критические значения для этой статистики были рассчитаны с помощью симуляций. Для заданной длины выборки *n* и значения ошибки первого рода α генерировались 100 тысяч выборок из независимых случайных величин, имеющих стандартное нормальное распределение. Выбор параметров распределения здесь не имеет значения, так как распределение t-статистики зависит лишь от суммарного числа наблюдений в двух выборках (в рассматриваемом случае это длина ряда, в котором ищется сдвиг). Для каждой сгенерированной выборки рассчитывались критические значения $T_{\max,n,\alpha}$. В качестве оценки для значения $T_{\max,n,\alpha}$ использовались выборочные квантили этой статистики порядка $1-\alpha$. Таким образом, если пренебречь погрешностью вычислений, можно утверждать, что при отсутствии сдвига величина $T_{\max,n}$ превышает критическое значение $T_{\max,n,\alpha}$ с вероятностью α, которую исследователь может выбирать по собственному желанию.

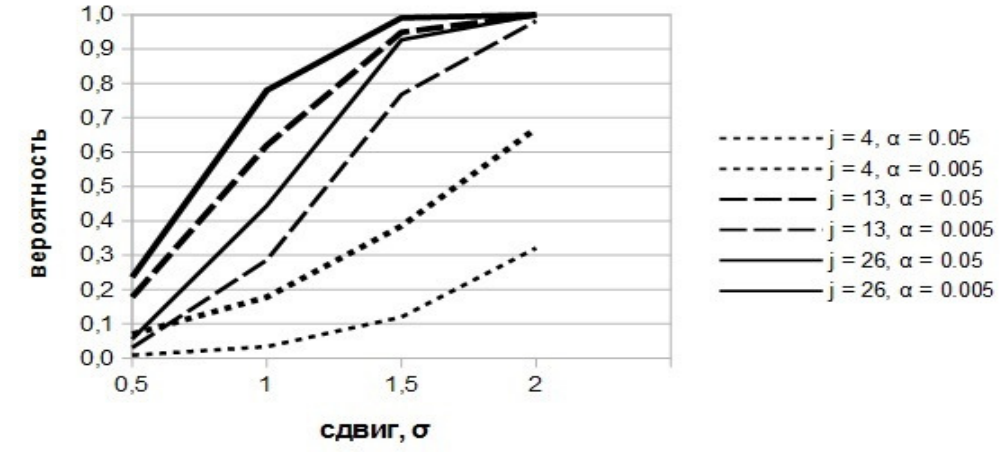
n \ α	0,05	0,02	0,01	0,005	0,002	0,001
20	3,28	3,73	4,08	4,39	4,84	5,15
40	3,18	3,55	3,81	4,09	4,44	4,66
50	3,16	3,51	3,78	4,03	4,33	4,52
100	3,16	3,48	3,72	3,95	4,21	4,45

Табл.1 Критические значения статистики $T_{\max,n}$ для разных объёмов выборки *n* и уровней значимости α.

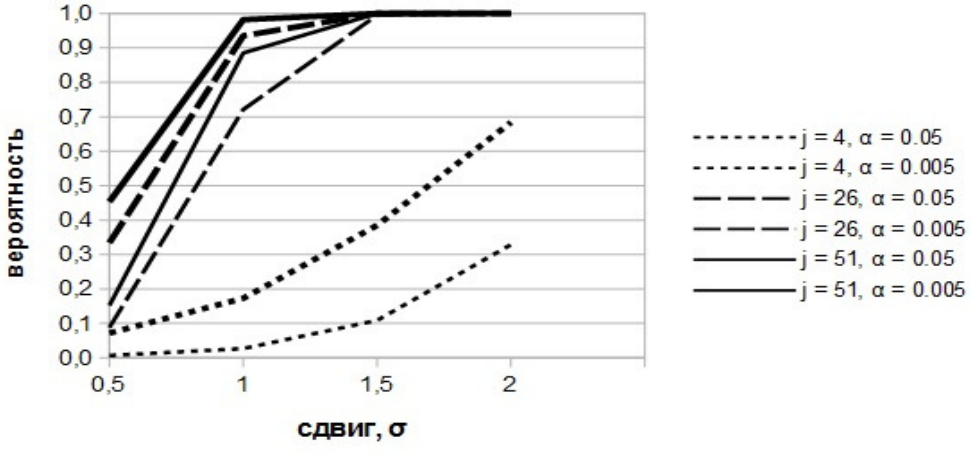
Рассчитанные критические значения статистики $T_{\max,n}$ для отдельных уровней значимости α и объёмов выборки n приведены в таблице 1.

Было замечено, что описанный тест не очень хорошо детектирует точки сдвига, лежащие близко к краям ряда. Также очевидно, что сдвиг тем труднее обнаружить, чем меньше величина, на которую меняется среднее. В связи с этим было проведено исследование частоты обнаружения точки сдвига в зависимости от её положения и величины сдвига.

Для различных комбинаций положения сдвига j и его величины генерировались случайные ряды длиной в 50 и 100 наблюдений, к которым применялся описанный выше тест. Как это часто делается при исследовании контрольных карт, величина сдвига задавалась в долях стандартного отклонения σ. Частота обнаружения сдвига при



(а)



(б)

Рис. 1. Вероятность обнаружения сдвига в зависимости от его величины, положения и выбранного уровня значимости для выборок объёма (а) 50 и (б) 100 наблюдений.

разных величинах уровня значимости α, подсчитанная на основании 10000 симуляций, отражена на рис. 1.

Приведенные данные показывают, что обнаружение точки сдвига вблизи начала ряда более трудно. Как видно из рис. 1, увеличение числа наблюдений не спасает положение — оценки мощности критерия для точки сдвига j=4 практически одинаковы для рядов в 50 и 100 наблюдений. Приведённые в приложении цифры говорят даже о небольшом снижении мощности при росте числа точек — впрочем, это может быть результатом погрешности вычислений. Если же сдвиг расположен в середине ряда или хотя бы удалён от края, вероятность обнаружения растёт с длиной ряда. Так, например, частота детектирования сдвига величиной 1σ в середине ряда длиной 50 при $\alpha = 0.05$ равна 0,78, а в середине ряда длиной 100 (при тех же параметрах) - 0,982.

j	α	Величина сдвига			
		$0{,}5\sigma$	1σ	$1{,}5\sigma$	2σ
4	0.05	0,070	0,178	0,385	0,669
4	0.005	0,009	0,034	0,120	0,321
13	0.05	0,178	0,620	0,949	0,999
13	0.005	0,031	0,286	0,768	0,982
26	0.05	0,236	0,780	0,991	1,000
26	0.005	0,056	0,444	0,927	0,999

Табл.2 Вероятность обнаружения точки сдвига для ряда длиной 50

j	α	Величина сдвига			
		$0{,}5\sigma$	1σ	$1{,}5\sigma$	2σ
4	0.05	0,072	0,173	0,385	0,684
4	0.005	0,007	0,027	0,109	0,329
26	0.05	0,334	0,935	1,000	1,000
26	0.005	0,088	0,721	0,995	1,000
51	0.05	0,453	0,982	1,000	1,000
51	0.005	0,152	0,885	1,000	1,000

Табл.3 Вероятность обнаружения точки сдвига для ряда длиной 100

3.2 Уменьшение количества ложных детектирований

Решение проблема ошибочных обнаружений сдвигов весьма актуально при работе с большими временными рядами. Здесь уместно привести следующий пример. Допустим, у нас есть ряд без сдвигов длиной $N = 10^6$. Разобьем его на сегменты длины $n = 100$, таких сегментов будет $s = 10^4$. Вероятность ошибки второго рода возьмем

$\alpha = 0{,}005$. Тогда количество ложно детектированных точек сдвига может достигать $s \cdot \alpha = 50$. Это число может быть еще больше из-за проверки пограничных областей между сегментами без сдвигов. Таким образом, представляется целесообразным верифицировать найденные с помощью статистики точки, используя дополнительную процедуру проверки.

В данной работе в качестве такой дополнительной проверки используется метод, основанный на идее информационного критерия (см. [14], [15]). Введем функцию стоимости (cost function) $C\left(\{x_i\}_{i=1}^{n}\right)$ временного ряда $\{x_i\}_{i=1}^{n}$. Точку τ будем считать точкой сдвига, если

$$C\left(\{x_i\}_{i=1}^{\tau}\right) + C\left(\{x_i\}_{i=\tau+1}^{n}\right) + \beta < C\left(\{x_i\}_{i=1}^{n}\right),$$

где β-некоторое штрафное слагаемое. В соответствии с информационным критерием Шварца[16] возьмем $\beta = \ln(n)$. Следуя принятой в литературе практике в качестве функции стоимости будем использовать функцию максимального правдоподобия.

Вычислительные эксперименты показывают, что если каждую точку, найденную статистикой проверять по описанной формуле, то количество ложных детектированний можно снизить в среднем на 35%

4. Поиск точек сдвига среднего временного ряда.

Предположим, что расстояние между точками сдвига не менее n и что временной ряд имеет длину $N = sn$. Пусть для ряда длины n вычислено граничное значение статистики $T_{\max,n}$ (п. 3.1). Разобьем ряд на сегменты длиной n. Для k-го ($1 \le k \le s$) сегмента с помощью описанного теста определим наличие (и при необходимости – местоположение) внутри данного сегмента точки сдвига. В случае обнаружения точки сдвига проверим ее с помощью подхода, описанного в 3.2.

Как уже говорилось, тест из п.3.1 не очень хорошо обнаруживает точки сдвига вблизи концов отрезка. Поэтому производится дополнительная проверка областей вокруг точек, разделяющих сегменты. Допустим, что для ряда длины l ($l < n$) тоже вычислено предельное значение статистики $T_{\max,n}$. Тогда, если на k м и $k+1$ м отрезке статистика не обнаружила точек сдвига – применим ее к отрезку ряда $\left\{x_i \middle| i = nk - l + 1..nk + l\right\}$. Эта дополнительная проверка эффективно находит точки сдвиги вблизи концов сегментов.

Данный метод позволяет найти не более $2s-1$ точек сдвига. Поэтому он является приближенным методом решения проблема поиска точек сдвига, как и широко используемый метод бинарной сегментации.

5. Параллельный алгоритм.

Благодаря разбиению ряда на сегменты равной длины, параллелизация описанного алгоритма представляет собой намного более простую задачу, чем распараллеливание популярного метода бинарной сегментации и методов, использующих динамическое программирование.

Очевидно, что каждому вычислительному узлу необходимо выдать порцию сегментов ряда. Напомним, что в случае если на двух сегментах не найдено точек сдвига, производится проверка окрестности границы между этими сегментами. Для обеспечения полной независимости обработки всех таких "пограничных областей" предлагается сделать детектор двухпроходным. На первом проходе производится параллельный поиск сдвигов на сегментах (включая дополнительную проверку с помощью критерия Шварца, см п.3.2).Каждый вычислительный узел обрабатывает свою порцию сегментов. На втором проходе каждый вычислительный узел обрабатывает свою порцию граничных точек(точек, разделяющих сегменты). Если оказывается, что на сегментах слева и справа от данной граничной точки не найдено точек сдвига, то производится проверка пограничной области, как это описано в п.4.

Ниже приводится запись описанного параллельного алгоритма (в варианте для машины с общей памяти) в псевдокоде.

```
parallel for i = 1:n_of_segments
   1)с помощью статистики проверить на i-м сегменте гипотезу H_0;
   2)в случае отклонения гипотезы проверить наличие точки сдвига с
     помощью критерия Шварца;
   3)при наличии точки сдвига found_chp[i] = 1, иначе found_chp[i] = 0;

parallel for i = 1:n_of_int_bnd
      если  found_chp[i] = found_chp[i+1] = 0
        проверить область вокруг i-й граничной точки
```

Здесь n_of_segments - число сегментов, n_of_int_bnd = n_of_segments - 1 - число внутренних граничных точек.

Параллельная реализация алгоритма была написана на языке C с использованием технологии OpenMP. Вычисления проводились на мультипроцессорном сервере с общей памятью IBM eServer p690 Regatta (16 процессоров POWER4 с частотой 1.3 ГГц, 64 Гб ОЗУ, пиковая производительность 83.2 Gflops), входящем в суперкомпьютерный комплекс МГУ.

Масштабируемость алгоритма проиллюстрируем следующим примером. Был сгенерирован временной ряд с тремя точками сдвига. Время работы нашего детектора на различном числе вычислительных

узлов приведено в табл.4 Видно, что ускорение близко к линейному. Отметим, что во всех экспериментов все три точки сдвига были найдены.

Количество узлов	Время работы детектора, сек
1	9.18064
2	4.61888
3	3.0974
4	2.33147
5	1.87753
6	1.57556
7	1.36112
8	1.20629

Табл.4 Время работы детектора на различном числе вычислительных узлов

Литература

1. *Ranganathan A.* PLISS: Detecting and Labeling Places Using Online Change-Point Detection Proceedings of Robotics: Science and Systems VI, 2010
2. *Lenardon M. J., Amirdjanova A.* Interaction between stock indices via changepoint analysis. Appl. Stochastic Models Bus. Ind., 2006, v. 22, pp 573-586.
3. *Braun J. V., Braun R. K., Muller H. G.* (2000). Multiple changepoint fitting via quasilikelihood, with application to DNA sequence segmentation. Biometrika, 87(2):301–314
4. *Shewhart W.* The application of statistics as an aid in maintaining quality of a manufactured product. Journal of the American Statistical Association, 1925, 20: 546–548.
5. *Page E.S.* Continuous Inspection Schemes. Biometrika, 1954, Vol. 41(1), pp. 100-115
6. *Shiryaev A. N.* The problem of the most rapid detection of a disturbance in a stationary process, Soviet Mathematics—Doklady 2, 1961: 795–799
7. *Roberts S. A* comparison of some control chart procedures. Technometrics 1966; 8 :411–430
8. *Basseville M., Nikiforov I.* Detection of Abrupt Changes: Theory and Application Prentice-Hall: Englewood Cliffs, NJ, 1993.
9. *Edwards AWF, Cavalli-Sforza L.L.* A Method for Cluster Analysis. Biometrics, 1965, 21(2), pp. 362–375.
10. *Killick R, Fearnhead P, Eckley I.A*. Optimal Detection of Changepoints with a Linear Computational Cost. Journal of the American Statistical Association, 2012, 107(500), 1590–1598.

11. *Jackson B., Sargle J. D., Barnes D., Arabhi S., Alt A., Gioumousis P., Gwin E., Sangtrakulcharoen P., Tan L., Tsai T. T.* (2005). An algorithm for optimal partitioning of data on an interval. IEEE, Signal Processing Letters, 12(2):105–108.
12. *Hawkins D.M., Peihua Qiu, Chang Wook Kang* The Changepoint Model for Statistical Process Control Journal of Quality Technology, 2003, Vol.35, Issue 4, p355-366
13. *Sen A., Srivastava M. S.* On Tests for Detecting Change in Mean When Variance is Unknown. Annals of the Institute of Statistical Mathematics, 1975, v.27, pp. 479–486.
14. *Akaike H.* Information theory and an extension of the maximum likelihood principle. S. Kotz, and N.L. Johnson (eds.) Breakthroughs in Statistics, 1992, Vol.1: 610–624. Springer-Verlag, London.
15. *Killick R, Eckley I.A.* Changepoint: An R Package for Changepoint Analysis. Journal of Statistical Software, Volume 58, Issue 3.
16. *Schwarz D.* Estimating the dimension of a model, Annals of Statistics, 6, 461-464