I. V. Toupitsyn

FAST ALGORITHM FOR RECONSTRUCTION OF INTERMEDIATE VIEWS FROM STEREOPAIR

An effective algorithm for stereoreconstruction of A. A. Lukianitsa is considered in this article. An algorithm modification, allowing to increase the speed of the algorithm is suggested. The results of carried out experiments are presented as well.

Keywords: stereo image, epipolar geometry, corresponding points, disparity.

УДК 330.43

С. В. Вохмянин

ИСПЫТАНИЕ АЛГОРИТМА МЕТОДА «ГУСЕНИЦА-SSA» ДЛЯ ВОССТАНОВЛЕНИЯ ВРЕМЕННОГО РЯДА

Рассмотрен базовый алгоритм метода «Гусеница-SSA» и проведены его испытания.

Ключевые слова: выделение тренда, нахождение периодик, устранение шума, разложение временного ряда на компоненты.

Одной из важнейших задач в анализе временных рядов является отделение тренда и периода от шума. Данная статья посвящена исследованию мощного и быстроразвивающегося метода анализа временных рядов «Гусеница-SSA» [1].

Рассмотрим временной ряд F:

$$f_0, f_1, ..., f_{N-1}, \quad (1)$$

где N – его длина. В дальнейшем будем предполагать, что ряд F – ненулевой.

Алгоритм метода «Гусеница-SSA» состоит из четырех последовательно выполняемых шагов: вложения, сингулярного разложения, группировки и диагонального усреднения.

На *первом шаге* процедура вложения переводит исходный временной ряд F в последовательность многомерных векторов, которая называется *траекторной матрицей*.

Для анализа временного ряда выбирается целочисленный параметр L, именуемый *длиной окна*, такой что $1 < L < N$. При этом образуется $K = N - L - 1$ векторов вложения:

$$X_i = (f_{i-1}, f_i, ..., f_{i+L-2})^T, \ 1 \le i \le K .$$

Эти векторы образуют траекторную матрицу ряда F, столбцами которой являются скользящие отрезки ряда длины L: с первой точки по L-ю, со второй по $(L + 1)$-ю и т. д.:

$$X = [X_1 : X_2 : ... : X_K] = \begin{pmatrix} f_0 & f_1 & ... & f_{K-1} \\ f_1 & f_2 & ... & f_K \\ ... & ... & ... & ... \\ f_{L-1} & f_L & ... & f_{N-1} \end{pmatrix}. \quad (2)$$

Существует взаимно однозначное соответствие между матрицами размерности $L \times K$ вида (2) и рядами (1) длины $N = L + K - 1$ [1].

Результатом *второго шага* является сингулярное разложение траекторной матрицы (2) в сумму элементарных матриц.

Пусть $S = X \cdot X^T$. Обозначим через $\lambda_1, \lambda_2, \ldots, \lambda_L$ собственные числа матрицы S, взятые в неубывающем порядке, а через $U_1, U_2, \ldots, U_L$ ортонормированную систему собственных векторов матрицы S, соответствующих упорядоченным собственным числам. Тогда сингулярное разложение траекторной матрицы X может быть записано следующим образом:

$$X = \sum V_i , \quad (3)$$

где $V_i = U_i \cdot U_i^T \cdot X$, $I = 1, \ldots, L$. Учитывая, что каждая из матриц V_i имеет ранг 1, назовем их элементарными матрицами [1].

Предположим, что исходный временной ряд является суммой нескольких рядов, что позволяет при некоторых условиях определить по виду собственных чисел и собственных векторов, какие это слагаемые и какой набор элементарных матриц соответствует каждому из них.

На *третьем шаге* на основе разложения (3) множество индексов $\{1, 2, \ldots, L\}$ делится на m непересекающихся подмножеств $I_1, I_2, \ldots, I_m$. Тем самым разложение (3) может быть записано в виде

$$X = \sum_{i=1}^{m} Y_i , \quad (4)$$

где $Y_i = \sum_{k \in I_i} V_k$ – результирующие матрицы для каждого подмножества I_i, $I = 1, \ldots, m$.

Фактически именно на этом шаге происходит разделение исходного ряда (1) на шумы, тренд и периодики. Основным критерием группировки является значимость

каждой элементарной матрицы V_k, прямо соответствующая ее собственному числу λ_k.

На *четвертом шаге* алгоритма каждая матрица сгруппированного разложения (4) переводится в ряд длины N.

Положим $L^* = \min(L, K)$, $K^* = \max(L, K)$. Пусть также $y^*_{ij} = Y_{ij}$, если $L < K$, и $y^*_{ij} = Y_{ji}$, если $L > K$. Диагональное усреднение переводит каждую результирующую матрицу $Y^{(s)}$, $s = 1, 2, \ldots, m$, в ряд $\tilde{f}^{(s)}$ по следующей формуле:

$$\tilde{f}_k = \begin{cases} \dfrac{1}{k+1}\sum\limits_{n=1}^{k+1} y^*_{n,k-n+2}, & 0 \le k < L^* - 1, \\ \dfrac{1}{L^*}\sum\limits_{n=1}^{L^*} y^*_{n,k-n+2}, & L^* - 1 \le k < K^*, \\ \dfrac{1}{N-k}\sum\limits_{n=k-K^*+2}^{N-K^*+1} y^*_{n,k-n+2}, & K^* \le k < N. \end{cases} \qquad (5)$$

Эта формула соответствует усреднению элементов вдоль диагоналей $I + j = k + 2$.

Итак, применяя диагональное усреднение (5) к результирующим матрицам $Y^{(s)}$, получаем ряды $\tilde{F}^{(s)} = (\tilde{f}_0^{(s)}, \tilde{f}_1^{(s)}, \ldots, \tilde{f}_{N-1}^{(s)})$. Исходный же ряд F раскладывается в сумму m рядов:

$$F = \sum_{n=1}^{m} \tilde{F}^{(s)}, \quad f_i = \sum_{n=1}^{m} \tilde{f}_i^{(s)},$$
$$n = 0, 1, \ldots, N-1, s = 1, 2, \ldots, m. \qquad (6)$$

Таким образом, результатом работы алгоритма является разложение временного ряда на интерпретируемые аддитивные составляющие. При этом он не требует стационарности ряда, знания модели тренда, а также сведений о наличии в ряде периодических составляющих и их периодах. При таких слабых предположениях метод «Гусеница-SSA» может решать различные задачи, такие как выделение тренда, обнаружение периодик, сглаживание ряда, построение полного разложения ряда в сумму тренда, периодик и шума [2].

Разумеется, данный метод имеет и свои недостатки. Во-первых, для получения составляющих исходного ряда используется неавтоматическая группировка компонент сингулярного разложения траекторной матрицы ряда (хотя залог успешного разложения заключается как раз в правильной группировке). Во-вторых, отсутствие модели не позволяет проверять гипотезы о наличии в ряде той или иной составляющей (этот недостаток объективно присущ всем непараметрическим методам). Отметим также, что рассматриваемый нами непараметрический метод в некоторых случаях позволяет получить результаты, часто незначительно менее точные, чем многие параметрические методы при анализе ряда с известной моделью [3].

Для исследования преимуществ и недостатков алгоритма метода «Гусеница-SSA» рассмотрим его работу на трех различных примерах. В каждом из примеров дан временной ряд, который состоит из суммы сгенерированных помех R_i и заданной искомой функции x_i:

$$f_i = x_i + R_i .$$

Введем также критерий эффективности, задаваемый отношением

$$W = \frac{\sum (A_i - x_i)^2}{\sum (R_i)^2} \cdot 100\ \%, \qquad (7)$$

где A_i – восстановленный (очищенный от помех) ряд, полученный с помощью алгоритма. В (7) числитель является суммой квадратов отклонений восстановленного ряда от чистого, в то время как знаменатель есть сумма квадратов помех. Таким образом, (7) показывает долю помех, не отделенную после применения алгоритма, поэтому будем называть его *гашением шума*.

Пример 1. Простой временной ряд, слабые помехи; $x_i = I + 10$, $I = 0, 1, \ldots, 49$, $N = 50$, $L = 25$; R_i – равномерно распределенная случайная величина из промежутка $[-2; 2]$. Матрица S имеет размеры 25×25 и 25 собственных чисел λ_i (табл. 1).

В качестве индексов группировки выбираются числа 24 и 25 как соответствующие наиболее значимым составляющим. С ними соотносятся элементарные матрицы V_{24} и V_{25}. Производя усреднение для результирующей матрицы $Y^0 = V_{24} + V_{25}$, получаем восстановленный ряд (рис. 1).

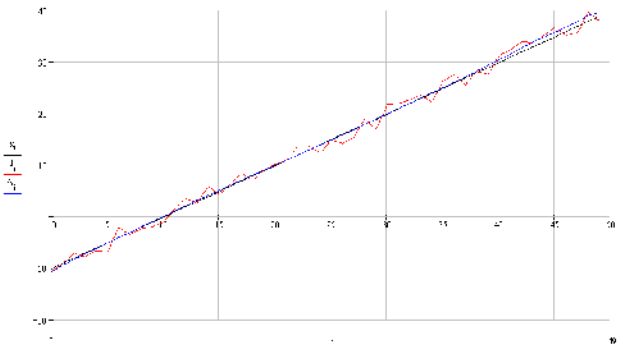

Рис. 1. Графики рядов (для примера 1): чистого, с шумом и восстановленного

Гашение шума составило $W = 11{,}4$ % от исходного шума.

Пример 2. Ряд с сезонными составляющими, средние помехи; $x_i = \dfrac{i(i-60)}{100} + 5\sin(i)$, $I = 0, 1, \ldots, 59$, $N = 60$, $L = 30$; R_i – равномерно распределенная случайная величина из промежутка $[-3; 3]$. Матрица S имеет размеры 30×30 и 30 собственных чисел λ_i (табл. 2).

В качестве индексов группировки выбираются числа с 27 по 30 как соответствующие наиболее значимым составляющим (стоит отметить, что они не обязаны быть последними, хотя такое довольно часто происходит). Им соответствуют элементарные матрицы V_{27}, V_{28}, V_{29} и V_{30}. Производя усреднение для результирующей матрицы $Y^0 = V_{27} + V_{28} + V_{29} + V_{30}$, получаем восстановленный ряд (рис. 2).

Гашение шума составило $W = 25{,}6$ % от исходного шума.

Пример 3. Ряд с несколькими сезонными составляющими, сильные помехи; $x_i = 0{,}03i + 1{,}6\sin(0{,}3i + 0{,}17) + 1{,}3\sin(2i + 0{,}57)$, $I = 0, 1, \ldots, 49$, $N = 50$, $L = 15$; R_i – нормально распределенная случайная величина, $\sigma = 3$. Матрица S имеет размеры 15×15 и 15 собственных чисел λ_i (табл. 3).

В данном случае из-за сильного шума выбор компонент для группировки довольно затруднителен и распознать тренд и периодики сложно. Анализ показал, что уве-

личение количества индексов в подобной ситуации приводит к тому, что восстанавливаются не только аддитивные компоненты, но и неотделяемый шум.

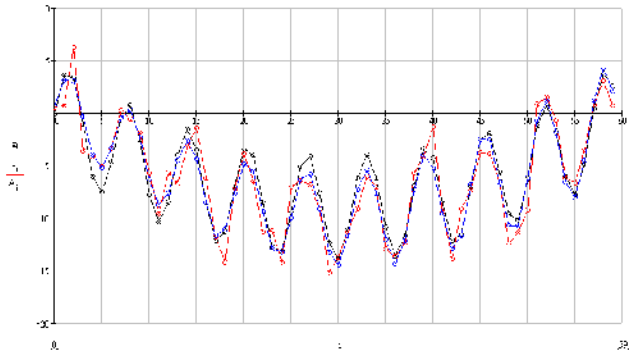

Рис. 2. Графики рядов (для примера 2): чистого, с шумом и восстановленного

Гашение шума при взятии трех наиболее значимых компонент составляет $W = 21{,}8\ \%$, для четырех – $W = 29{,}2\ \%$ и для пяти – $W = 34{,}6\ \%$.

Результаты для трех выделенных компонент приведены ниже (рис. 3).

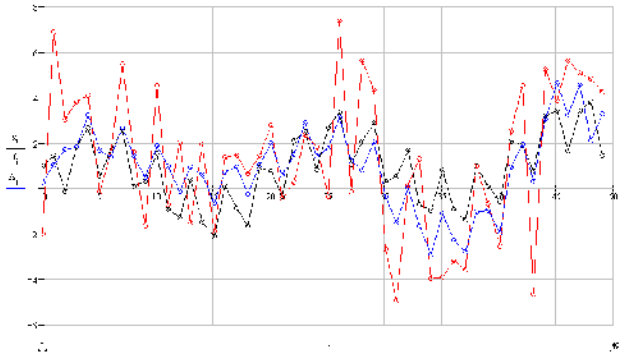

Рис. 3. Графики рядов (для примера 3): чистого, с шумом и восстановленного

По результатам проведенных испытаний можно сделать вывод, что базовый алгоритм метода «Гусеница-SSA» справляется с поставленной для него задачей: для временного ряда отделяет тренд и периодики от помех, снижая уровень шумов в 2–3 раза; при этом изначально неизвестно, какой тип будут иметь значимые компоненты: линейный, периодический, логарифмический или иной. Это является достоинством данного метода, что в перспективе позволит создать мощный механизм непараметрического анализа временных рядов, в том числе и в виде программ для ЭВМ.

Недостатками же базового алгоритма метода «Гусеница-SSA» являются необходимость вмешательства человека для анализа разделенных компонент и проблема выбора длины окна, от которой зависит качество разделения аддитивных составляющих. Дальнейшие исследования будут направлены на автоматизацию процесса анализа и использование других методов, улучшающих качество результатов работы алгоритма и уменьшающих вмешательство человека в этот процесс.

Библиографические ссылки

1. Голяндина Н. Э. Метод «Гусеница-SSA»: анализ временных рядов : учеб. пособие. СПб., 2004.

2. Главные компоненты временных рядов: метод «Гусеница» / под ред. Д. Л. Данилова, А. А. Жиглявского. СПб. : Пресском, 1997.

3. Golyandina N., Nekrutkin V., Zhigljavsky A. Analysis of Time Series Structure: SSA and Related Techniques. London : Chapman& Hall/CRC, 2001.

Таблица 1

Вклад собственных чисел λ_i матрицы *S*, в процентах от их суммы, для примера 1

i	1	2	3	4	5	6	7	8	9	10	11	12	13
λ_i, %	0,00	0,00	0,00	0,00	0,00	0,00	0,01	0,01	0,01	0,00	0,01	0,01	0,01
i	14	15	16	17	18	19	20	21	22	23	24	25	–
λ_i, %	0,02	0,02	0,02	0,02	0,03	0,03	0,03	0,04	0,08	0,08	2,76	96,8	–

Таблица 2

Вклад собственных чисел λ_i матрицы *S*, в процентах от их суммы, для примера 2

i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
λ_i, %	0,03	0,03	0,03	0,03	0,04	0,05	0,02	0,07	0,07	0,07	0,08	0,01	0,00	0,00	0,12
i	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
λ_i, %	0,11	0,12	0,12	0,17	0,20	0,21	0,21	0,22	0,26	0,29	0,33	4,14	5,97	7,58	79,44

Таблица 3

Вклад собственных чисел λ_i матрицы *S*, в процентах от их суммы, для примера 3

i	1	2	3	4	5	6	7	8
λ_i, %	2,50	2,49	2,85	2,14	1,94	3,29	4,00	4,73
i	9	10	11	12	13	14	15	–
λ_i, %	5,26	6,83	7,78	9,99	11,31	13,83	21,07	–

S. V. Vohmyanin

TESTING THE ALGORYTHM OF THE METHOD «CATERPILLAR-SSA» FOR REESTABLISHING OF TIME SERIES

The base algorithm of the «Caterpillar-SSA» method is considered and tested.

Keywords: trend allocation, finding of the periodicals, noise elimination, decomposition of time series to components.

УДК 681.3

Е. А. Энгель

ИСПОЛЬЗОВАНИЕ ИНТЕЛЛЕКТУАЛЬНЫХ АЛГОРИТМОВ ДЛЯ ОБРАБОТКИ ТЕКСТОВОЙ ИНФОРМАЦИИ*

С целью создания программной системы для адаптивного текстового реферирования разработаны и реализованы в виде отдельного модуля интеллектуальные алгоритмы автоматического определения жанра текста. Модуль позволяет нормализовать 45 статистических параметров: лексических, синтаксических, позиционных и дискурсивных; группировать гетерогенные параметры с помощью алгоритма K-средних; выполнять факторный анализ; ранжировать параметры, существенные для идентификации научного жанра, публицистики и беллетристики, посредством двух алгоритмов.

Ключевые слова: обработка текстовой информации, интеллектуальные алгоритмы, алгоритм K-средних.

В течение прошлого десятилетия автоматическое определение жанра текста стало важной проблемой, исследованной в пределах такой научной области, как обработка естественного языка. Будучи интересной с теоретической точки зрения, задача определения жанра тесно связана с развитием информационного поиска цифровых библиотек и реферирования. Автоматическую идентификацию интернет-жанров можно считать отдельной предметной областью, которая обращается к реальной проблеме информационной перегрузки и играет существенную роль в улучшении часто неадекватных результатов работы поисковых машин.

Включение модуля автоматического определения жанра текста позволяет оптимизировать и повысить эффективность системы текстового реферирования. Стимулом для создания такого модуля стали результаты оценки эффективности следующих систем автоматического реферирования: Event Tracking Summarizer, Subject Search Summarizer, Copernic Summarizer и Open Text Summarizer. Программный продукт Event Tracking Summarizer, специально разработанный для обработки беллетристики, оказался эффективнее других систем автоматического реферирования в среднем на 15 % для беллетристики и менее эффективным для других жанров. Следовательно, возникает необходимость в создании адаптивной системы текстового реферирования на основе алгоритмов, оптимизированных для конкретного текстового жанра.

У любой NLP-системы есть модуль предварительной обработки, который в зависимости от текстовых задач обработки выполняет лексическое и синтаксическое разложение, стемминг, аннотацию и синтаксический парсинг. Результатом предварительной обработки является модель объекта, которая отражает лингвистические характеристики входного текста, например слов, фраз, предложений, параграфов. Далее лингвистические характеристики ранжируются, в результате чего получается список текстовых параметров. Параметры с самыми высокими весами затем сравниваются с эталонными моделями, хранящимися в лингвистической базе данных. Входной текстовый жанр идентифицируется в зависимости от степени соответствия между распределением параметров в этом тексте и в одной из эталонных моделей. На следующей стадии система применяет алгоритмы реферирования, оптимизированные для данного жанра (рис. 1).

Предметом данной статьи является задача определения жанра текста; алгоритмы реферирования выходят за рамки данной публикации.

Параметры, идентифицирующие жанр. Идентификация жанра текста основана на анализе набора параметров, являющихся лингвистическими признаками с назначенными весами, т. е. некоторыми числовыми значениями, отражающими его важность для данного текста. Следовательно, процесс идентификации жанра включает две

* Работа выполнена в рамках Федеральной целевой программы «Научные и научно-педагогические кадры инновационной России на 2009–2013 гг.», Госконтракт 02.740.11.0663.