Вычислительные технологии, 2021, том 26, № 5, с. 95–105. © ФИЦ ИВТ, 2021
Computational Technologies, 2021, vol. 26, no. 5, pp. 95–105. © FRC ICT, 2021
ISSN 1560-7534
eISSN 2313-691X

ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ

DOI:10.25743/ICT.2021.26.5.008

Влияние мощности алфавита на качество восстановления символьной периодической последовательности по последовательности с шумом

Г. Н. Жукова[1,*], М. В. Ульянов[2,3]

[1]Национальный исследовательский университет "Высшая школа экономики", 109028, Москва, Россия

[2]Институт проблем управления им. В.А. Трапезникова РАН, 117997, Москва, Россия

[3]Московский государственный университет им. М.В. Ломоносова, 119991, Москва, Россия

*Контактный автор: Жукова Галина Николаевна, e-mail: gzhukova@hse.ru

Поступила 6 августа 2021 г., доработана 12 августа 2021 г., принята в печать 18 августа 2021 г.

В статье рассмотрена задача восстановления символьных периодических последовательностей, искаженных шумами вставки, а также замены и удаления символов. Поскольку степень детализации символьного описания процесса определяется мощностью алфавита, представляет интерес исследование влияния степени детализации символьного описания на возможность восстановления полной информации об исходной периодической последовательности. Представлено экспериментальное исследование зависимости характеристик качества предложенного авторами метода восстановления периода от мощности алфавита. Для алфавитов разной мощности приводятся доля последовательностей с удовлетворительно восстановленным периодом и относительная погрешность определения длины периода. Качество восстановления оценивается отношением редакционного расстояния от восстановленной периодической последовательности до исходной строго периодической последовательности.

Ключевые слова: символьная последовательность, мощность алфавита, периодическая последовательность, зашумленная последовательность, шум вставки, шум удаления, шум замены.

Цитирование: Жукова Г.Н., Ульянов М.В. Влияние мощности алфавита на качество восстановления символьной периодической последовательности по последовательности с шумом. Вычислительные технологии. 2021; 26(5):95–105. DOI:10.25743/ICT.2021.26.5.008.

Введение

Актуальность данного исследования связана с наличием широкого круга прикладных задач обработки и анализа данных реального мира, в которых информацию целесообразно кодировать символами конечного алфавита. Изучение особенностей качественного поведения исследуемых объектов или процессов относится к задачам качественного анализа. Их решение основано на предобработке числовых данных о наблюдаемых состояниях исследуемого процесса символьным кодированием в конечном алфавите. Преимущества такого подхода обусловлены возможностью отбросить несущественные детали, которые не несут полезной информации в аспекте задач качественного анализа.

Аппарат символьного кодирования позволяет путем варьирования мощности алфавита выбирать степень детализации описания процесса, достаточную для решения задачи анализа. Символьное кодирование актуально также при качественном анализе больших данных, поскольку высокая точность числовых представлений признаков приводит к неоправданно большим объемам информации и трудоемким вычислениям без улучшения качества получаемых результатов [1, 2].

Периодические процессы широко представлены в таких предметных областях, как биоинформатика, медицина, цифровая экономика, прогнозирование временны́х рядов, анализ бизнес-процессов, при символьном кодировании им соответствуют периодические символьные последовательности — слова над конечным алфавитом. Методы поиска периодичности для таких последовательностей достаточно хорошо разработаны [3].

Однако для ряда предметных областей, в которых возможно применение символьного кодирования траекторий изучаемых процессов, исследователи сталкиваются с искажениями, фрагментарностью информации и наличием шумов. В реальности наблюдаемые значения исследуемых процессов, представляемые в виде элементов временны́х рядов, содержат как ошибки измерения, так и случайные искажения, вызванные внешними факторами. Такого рода ошибки в общем случае трактуются как шум. При исследовании временны́х рядов и их прогнозировании работа с зашумленными данными вызывает значительные трудности, что приводит к формулировке задачи устранения шума.

Для шумоподавления в числовых данных используются методы скользящего среднего, экспоненциального сглаживания и др. [4]. Однако эти методы не применимы при работе с зашумленными символьными последовательностями. В этих условиях при решении задачи поиска периодичности, включающей в себя как определение периодически повторяющегося символьного фрагмента, так и его длины, далее называемой периодом, возникает задача уменьшения влияния шума на результаты исследования. Один из возможных подходов к определению символьной периодичности в зашумленных словах описан в [5, 6].

В данной статье предполагается, что в результате измерений или эксперимента исследователь получает не строго периодическую символьную последовательность, а последовательность, искаженную шумами, имеющими различную природу. Предполагается также, что уровень внесенного шума не превышает 10 %. Под уровнем шума понимается отношение редакционного расстояния между искаженной и исходной последовательностями к длине последней, выраженное в процентах. Восстановление проводится методом, предложенным авторами в [5, 6], где рассмотрен случай только бинарного алфавита. Поскольку степень детализации символьного описания процесса определяется мощностью алфавита, представляет интерес исследование влияния степени такой детализации на возможность восстановления полной информации об исходно периодических словах.

В настоящей статье изучается поведение характеристик качества восстановления периодической последовательности методом из [5] в зависимости от мощности алфавита в диапазоне от четырех до шестидесяти.

1. Постановка задачи

При изучении последовательностей над бинарным алфавитом [5] мы рассматриваем задачу построения периодической последовательности, содержащей не менее восьми пери-

одически повторяющихся фрагментов (из которых последний может быть неполным), на основе заданной последовательности, полученной внесением шумов вставки, удаления и замены в некоторую строго периодическую последовательность. Символьную последовательность $q^\sigma = s_1, s_2, \ldots, s_n$ над некоторым конечным алфавитом Σ ($\sigma = |\Sigma| \geq 2$, где s — произвольный символ из Σ) будем называть словом длины n. Полагаем, что конечный алфавит Σ мощности $\sigma = |\Sigma|$ изначально известен. Любую последовательность символов $s_k, s_{k+1}, \ldots, s_{m-1}, s_m$, $1 \leq k \leq m \leq n$, будем называть подсловом (фрагментом) слова q^σ.

Для исследования влияния мощности алфавита на качество распознавания периодической последовательности будем рассматривать в алфавитах различной мощности σ строго периодические слова $q^\sigma(m,p)$, содержащие одинаковое число $m \geq 8$ периодических повторяющихся фрагментов длины p. В каждое из слов $q^\sigma(m,p)$ по методике, предложенной в [7], вносятся шумы вставки, удаления или замены символов из того же самого алфавита Σ. В результате получаем слово $\widetilde{q}^\sigma = \widetilde{q}^\sigma(q^\sigma(m,p))$ длины n. Далее на основе слова $\widetilde{q}^\sigma$, используя предложенный в [5] алгоритм, пытаемся построить, если это возможно, слово $\bar{q}^\sigma = \bar{q}^\sigma(\widetilde{q}^\sigma(q^\sigma(m,p)))$ такой же длины n, как и $\widetilde{q}^\sigma$, аппроксимирующее исходное "неизвестное" (для алгоритма восстановления) слово $q^\sigma(m,p)$.

2. Методика оценки точности и качества восстановления периода

Точность δ определения периода $\bar{p} = \bar{p}(\bar{q}^\sigma)$, т. е. длины минимального периодически повторяющегося фрагмента, будем оценивать отношением

$$\delta = \frac{|p - \bar{p}|}{p},$$

где p — период исходной строго периодической последовательности $q^\sigma(m,p)$, не искаженный шумом; $\bar{p}$ — период, полученный методом из [5]. Точность δ вычисляется для каждой последовательности из серии зашумленных последовательностей, полученных на основе одной и той же исходной чисто периодической последовательности.

Качество восстановления периодической последовательности можно оценить с помощью отношения редакционного расстояния $d(\bar{q}^\sigma, q^\sigma(m,p))$ (по Левенштейну [8]) между восстановленной и исходной последовательностями к длине последней (равной mp, где m — число периодически повторяющихся фрагментов, p — период исходной последовательности):

$$\varepsilon(\bar{q}^\sigma, q^\sigma) = \frac{d(\bar{q}^\sigma, q^\sigma(m,p))}{mp}.$$

Иногда алгоритм находит период в два–три раза больше исходного, при этом редакционное расстояние между исходной последовательностью и частью восстановленной последовательности равной длины мало́. Поэтому для более адекватного вычисления ε сравниваются последовательности, построенные из повторяющихся фрагментов соответственно исходной и восстановленной последовательностей, имеющих длину, равную длине зашумленной последовательности. Таким образом, далее в статье качество восстановления оценивается как

$$\varepsilon_n(\bar{q}_n^\sigma, q_n^\sigma) = \frac{d(\bar{q}_n^\sigma, q_n^\sigma(m,p))}{n},$$

где n — длина зашумленной последовательности. Кроме того, вычисляются аналогичные характеристики сравнения для периодической и зашумленной, а также восстановленной и зашумленной последовательностей. Первая позволяет судить о том, насколько близок уровень шума, вычисленный на основе редакционного расстояния, к заданному уровню. Вторая показывает, насколько построенная периодическая аппроксимация близка к зашумленной последовательности, на основе которой проводилось восстановление периодического прототипа.

Для каждого эксперимента по восстановлению периода и периодического фрагмента по отдельной зашумленной последовательности вычисляется значение ε, а для совокупного анализа серии из ста экспериментов используются минимальное, максимальное, среднее значения и медиана полученной выборки.

Поскольку предложенный в [5] алгоритм в некоторых случаях не может построить периодическую последовательность, имеющую не менее восьми периодов, в серии экспериментов с фиксированными значениями параметров m, p, σ и уровнями шумов удаления, замены и вставки ω_{del}, ω_{ch}, ω_{ins} способность алгоритма находить аппроксимирующую периодическую последовательность будем оценивать такими величинами, как:

1) ω_0 — доля экспериментов, в которых решение найдено;
2) ω_1 — доля экспериментов, в которых период найден точно, т. е. $\bar{p} = p$;
3) ω_2 — доля экспериментов, в которых $\delta \leq 0.1$.

Доля экспериментов вычисляется как отношение числа экспериментов, удовлетворяющих условию, к общему числу экспериментов. Для удобства далее мы выражаем долю в процентах.

3. Схема вычислительного эксперимента

Основная идея эксперимента опирается на требование сохранения количества информации в исследуемых словах при различных алфавитах. Поскольку целью исследования является изучение влияния мощности алфавита на качество восстановления периодической последовательности, рассматриваем две группы алфавитов — алфавиты мощности $\sigma = 4, 6, \ldots, 14, 16$ и алфавиты мощности $\sigma = 10, 20, \ldots, 50, 60$ с кодированием без потери информации. Перекрытие групп по алфавиту мощности 10 выбрано с целью сравнительного анализа.

Схема эксперимента имеет следующий вид. Обозначим β_1 максимальную мощность (16) алфавита первой группы и β_2 — максимальную мощность (60) алфавита первой группы. Далее для алфавита мощности β_1 фиксируем длину фрагмента (период), образующего периодическую последовательность (слово) в β_1 символов. При этом сам фрагмент состоит из всех β_1 символов алфавита. Генерируем слово $q^{\beta_1}(8, \beta_1)$.

Поскольку для алфавита мощности β_1 порождающий фрагмент состоит из всех символов алфавита, для алфавитов меньшей мощности в пределах рассматриваемой группы, исходя из требования сохранения количества информации, кодируем этот фрагмент символами текущего алфавита следующим образом: рассматриваем исходный порождающий фрагмент как число в системе счисления по основанию β_1 и переводим его в систему счисления по основанию текущей мощности алфавита σ. Очевидно, с уменьшением мощности алфавита длина порождающего фрагмента p будет возрастать приблизительно в $\log_\sigma \beta_1$ раз. Получаем слова $q^\sigma(8, p)$ для $\sigma = 4, 6, \ldots, \beta_1 - 1$. Для второй

группы алфавитов — аналогично, начиная с периода β_2, состоящего из всех β_2 символов алфавита.

Для каждой такой последовательности $q^\sigma(8,p)$ и фиксированных значений уровней шумов удаления, замены и вставки ω_{del}, ω_{ch}, ω_{ins} получено 100 искаженных последовательностей по методике, предложенной нами в [7]. Вначале вносился шум удаления, затем — замены и в конце — шум вставки, при этом для каждого вида шума с помощью генератора псевдослучайных чисел были получены номера символов, которые нужно удалить или изменить или после которых нужно вставить новый символ. Номера символов, необходимые для внесения шума, получены с помощью случайного перемешивания целых чисел от 0 до $kp-1$, осуществляемого посредством numpy.random.shuffle из пакета numpy. Из перемешанного списка номеров выбиралось необходимое количество первых элементов списка. Количество номеров символов для внесения шума вычислялось так: $n_{type} = \omega_{type} pk$, $type = del, ch, ins$, где p — период, m — число периодов, del — удаление, ch — замена, ins — вставка. Получаем слова $\widetilde{q}^\sigma = \widetilde{q}^\sigma(q^\sigma(m,p))$.

После внесения шумов с использованием программной реализации алгоритма [5] исследовалась возможность построить, если это возможно, слово $\bar{q}^\sigma = \bar{q}^\sigma(\widetilde{q}^\sigma(q^\sigma(m,p)))$, т. е. проводилась оценка $\bar{p}$ периода p с использованием информации о частоте встречаемости подслов длины 10 [5]. В случае получения оценки $\bar{p}$ проводилось построение аппроксимирующего периодически повторяющегося фрагмента путем разбиения искаженной последовательности на последовательные подслова длины $\bar{p}$ (последнее подслово длины меньше $\bar{p}$ не учитывалось) и выбора такого из них, у которого редакционное расстояние до одного из оставшихся подслов минимально. В случае если таких подслов несколько, выбиралось первое подслово с минимальным редакционным расстоянием до другого подслова.

Далее вычислялись характеристии качества восстановления периодической последовательности δ и ε. Для каждой серии из ста экспериментов с одинаковыми значениями всех параметров вычислялись доли ω_i, $i = 0, 1, 2$, экспериментов, в которых:

а) решение найдено (ω_0);

б) период найден точно, т. е. $\bar{p} = p$, (ω_1);

в) $\delta \le 0.1$ (ω_2).

Эксперименты для слова $q^\sigma(8,p)$ повторялись с другими фиксированными значениями уровней шумов удаления, замены и вставки ω_{del}, ω_{ch}, ω_{ins}.

4. Результаты вычислительного эксперимента

На рисунке представлена зависимость доли экспериментов с точно определенным периодом от мощности алфавита, а также от уровня и структуры шума. В табл. 1 и 2 приведены значения ε_n. Рассмотрены два случая, в первом периодический фрагмент состоял из $\beta_1 = 16$ неповторяющихся символов алфавита мощности 16 с последующим перекодированием в алфавиты меньшей мощности, во втором — $\beta_2 = 60$.

На рисунке по горизонтальной оси отмечены мощности алфавитов σ, по вертикальной — доля ω_1 (в процентах) зашумленных последовательностей, для которых алгоритм точно определил период (длину минимального повторяющегося фрагмента), т. е. $\bar{p} = p$. В легендах рисунка уровни шума в процентах указаны в кортежах в следующем порядке: шум удаления, шум замены, шум вставки. Для рисунков, стоящих в одной строке, легенда приведена общая; так, на рисунках *а* и *б* изображена доля ω_1 в случае общего уровня шума 5 %, шумам одинаковой структуры соответствуют линии одного цвета на

Т а б л и ц а 1. Оценка ε_n качества восстановления последовательности при равномерном шуме мощности от 4 до 16

Table 1. Estimation ε_n of the quality of the sequence restoration, uniform noise, cardinality from 4 to 16

Шум, %	σ	ε_n, восстановленная и зашумленная				ε_n, чистая и восстановленная	
		мин.	макс.	среднее	медиана	макс.	среднее
3	4	2.6	2.9	2.9	2.9	0.2	0.01
	6	2.5	3.0	3.0	3.0	0.3	0
	8	2.7	3.0	3.0	3.0	0.0	0
	10	2.5	2.8	2.8	2.8	0.3	0.01
	12	2.8	3.1	3.1	3.1	0.4	0
	14	2.6	2.8	2.8	2.8	0.4	0
	16	2.7	2.9	2.9	2.9	0.4	0.01
4.5	4	4.1	4.4	4.3	4.3	0.2	0.01
	6	4.0	4.5	4.4	4.5	0.5	0
	8	4.3	4.7	4.6	4.7	0.3	0.01
	10	4.2	4.7	4.6	4.7	0.3	0
	12	4.3	4.7	4.6	4.7	0.7	0.02
	14	3.9	4.4	4.4	4.4	0.4	0
	16	4.1	4.7	4.6	4.7	0.4	0
6	4	5.3	5.9	5.7	5.8	0.2	0.01
	6	5.4	6.0	5.9	5.9	0.3	0.01
	8	5.4	6.0	5.9	6.0	0.3	0.01
	10	5.8	8.1	6.0	6.1	2.5	0.03
	12	5.6	8.3	6.2	6.3	2.8	0.04
	14	5.7	6.1	6.0	6.1	0.0	0
	16	5.3	16.4	5.9	5.9	11.9	0.12
7.5	4	7.0	10.8	7.7	7.4	4.7	0.36
	6	6.6	7.5	7.3	7.4	0.3	0
	8	7.0	11.1	7.6	7.5	4.6	0.05
	10	6.9	9.4	7.4	7.3	2.5	0.06
	12	6.6	9.7	7.2	7.3	2.8	0.05
	14	7.2	7.7	7.6	7.5	0.4	0.01
	16	7.0	9.8	7.6	7.6	3.1	0.07
9	4	8.1	13.9	9.1	8.8	7.8	0.53
	6	8.3	11.9	8.8	8.8	4.3	0.21
	8	8.2	12.8	8.9	8.8	4.6	0.14
	10	8.1	12.8	8.8	8.8	5.0	0.15
	12	8.2	13.2	8.9	8.7	5.6	0.23
	14	8.1	12.7	8.8	8.6	5.9	0.15
	16	8.2	11.3	8.7	8.8	3.1	0.07
10.5	4	9.6	14.6	11.1	10.2	6.3	1.38
	6	9.1	14.5	10.4	10.1	6.0	0.52
	8	9.9	16.1	10.9	10.4	6.8	0.68
	10	9.4	14.4	10.2	10.2	5.2	0.21
	12	9.4	16.5	10.5	10.2	7.6	0.49
	14	9.7	16.9	10.7	10.3	10.1	0.64
	16	10.0	16.8	10.9	10.4	9.0	0.82

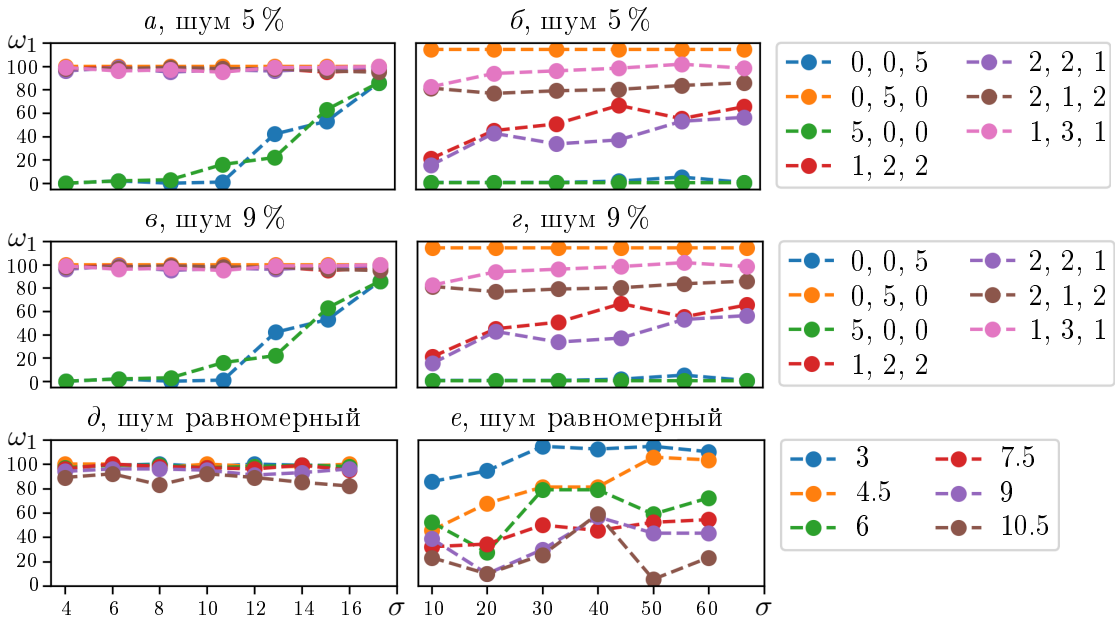
Т а б л и ц а 2. Оценка ε_n качества восстановления последовательности при равномерном шуме мощности от 10 до 60

Table 2. Estimation ε_n of the quality of the sequence restoration, uniform noise, cardinality from 10 to 60

Шум, %	σ	ε_n, восстановленная и зашумленная				ε_n, чистая и восстановленная			
		мин.	макс.	среднее	медиана	мин.	макс.	среднее	медиана
3	10	3	6	4	4	0.0	3	1.4	1.9
	20	3	5	3	3	0.0	2	0.7	0.0
	30	3	5	3	3	0.0	3	0.3	0.0
	40	3	6	4	3	0.0	3	0.6	0.0
	50	3	6	3	3	0.0	3	0.3	0.0
	60	3	6	3	3	0.0	3	0.4	0.0
4.5	10	4	7	6	6	0.0	4	2.1	1.9
	20	4	7	6	5	0.0	4	1.6	1.2
	30	4	7	6	6	0.0	4	1.5	1.4
	40	4	7	6	6	0.0	4	1.4	1.5
	50	4	7	5	4	0.0	3	1.0	0.0
	60	4	8	5	6	0.0	3	1.2	1.7
6	10	6	10	8	8	0.0	6	3.2	3.7
	20	7	11	9	9	1.2	6	3.3	3.6
	30	6	11	8	8	0.0	5	2.7	2.7
	40	6	10	8	8	0.0	4	2.5	3.0
	50	6	10	8	9	0.0	5	2.8	3.2
	60	6	10	8	8	0.0	5	2.5	3.3
7.5	10	7	13	11	11	0.0	7	4.4	4.7
	20	7	13	11	11	0.0	7	4.0	4.2
	30	7	13	10	10	0.0	7	3.9	4.1
	40	7	13	10	10	0.0	6	3.9	4.5
	50	7	13	10	10	0.0	7	3.5	3.2
	60	7	13	10	10	0.0	7	3.7	3.3
9	10	10	15	13	13	1.9	9	5.5	5.6
	20	10	15	13	13	1.2	8	5.2	4.8
	30	10	15	13	13	1.4	8	5.3	5.5
	40	10	16	12	12	1.5	9	4.8	4.5
	50	9	15	13	13	0.0	8	4.9	4.8
	60	9	16	12	13	0.0	8	4.6	5.0
10.5	10	12	17	15	15	2.8	10	6.6	6.5
	20	12	17	15	15	3.6	10	6.5	7.0
	30	13	18	15	15	2.7	11	6.3	6.1
	40	12	17	14	14	3.0	9	6.0	6.0
	50	12	17	15	15	2.9	10	5.9	6.4
	60	10	18	15	15	0.0	10	5.9	6.7

каждом из этих рисунков. Мы приводим результаты как для различных соотношений уровней шумов разных типов, так и для равномерного шума, когда внесенные шумы всех трех типов имеют одинаковый уровень.

В табл 1 и 2 представлены минимальное, максимальное, среднее и медианное значения отношения редакционного расстояния между восстановленной и зашумленной



Зависимость доли точно восстановленных периодов от мощности алфавита и от шума
Dependence of the fraction of accurately restored periods on the alphabet cardinality and noise

последовательностями, т. е. значение $\varepsilon_n(\bar{q}_n^\sigma, \tilde{q}_n^\sigma)$. Заметим, что последовательности выравниваются по длине зашумленной последовательности, что дает адекватный результат в случае, если найденный период в два или три раза больше исходного.

Также в табл. 2 приводятся минимальное, максимальное, среднее и медианное значения $\varepsilon_n(q_n^\sigma, \bar{q}_n^\sigma)$ для исходной строго периодической и восстановленной последовательностей с выравниванием по длине зашумленной последовательности, которое используется потому, что в случае применения алгоритма к реальным данным в распоряжении исследователя будет только зашумленная последовательность. В табл. 1 приводятся только максимальное и среднее значения $\varepsilon_n(q_n^\sigma, \bar{q}_n^\sigma)$; минимальное значение $\varepsilon_n(q_n^\sigma, \bar{q}_n^\sigma)$ и медиана в серии ста экспериментов не приводятся, потому что они равны нулю для всех случаев.

Данные приведены для равномерного шума, т. е. уровень шума каждого вида одинаков в каждом эксперименте и принимает значения от 1.0 до 3.5 % с шагом 0.5 %, так что общий уровень шума варьируется от 3.0 до 10.5 % с шагом 1.5 %.

5. Результаты и обсуждение

Вычислительный эксперимент подтвердил гипотезу о том, что структура шума заметно влияет на качество восстановления периодической последовательности. Причина в том, что при заметной разнице уровней шумов вставки и удаления существенно изменяется длина повторяющегося фрагмента (период), что, в свою очередь, ухудшает качество восстановления. В связи с этим при фиксированном общем уровне шума 5 и 9 % рассматривались предельные случаи, когда весь вносимый шум был только шумом замены, или только шумом вставки, или только удаления, а также случай равномерного шума (3, 3, 3) и некоторые другие.

На рисунке видно, что худшее качество восстановления наблюдается при заметном различии уровней шума удаления и вставки, а лучшее — при наличии только шума

замены. Так, при наличии только 5 %-ного шума замены мы наблюдаем 100 %-ную эффективность предложенного алгоритма во всем диапазоне мощности алфавитов — от 4 до 60. При одинаковых уровнях шума удаления и вставки качество определения периода хорошее, причем немного улучшается с ростом мощности алфавита для диапазона мощностей 10–60.

Еще при разработке метода восстановления периода на основе анализа подслов в скользящем окне авторами выдвинута гипотеза о том, что с увеличением мощности алфавита улучшается качество восстановления периодической последовательности предложенным в [5] алгоритмом. Эта гипотеза подтверждается нашими экспериментальными данными лишь в некоторых случаях, например при малых уровнях шума. Так, из данных рисунка *а* следует, что качество распознавания улучшается с ростом мощности алфавита только для худших случаев наличия только шума удаления или только вставки в первой группе мощностей алфавита (4–16). Данные рисунка *б* также свидетельствуют о положительной динамике качества восстановления для мощностей второй группы только при различии уровней шума вставки и удаления (шумы (1, 2, 2) и (2, 2, 1)). Заметим, что при общем уровне шума 9 % (рисунки *в* и *г*) гипотеза уже не подтверждается.

Можно сделать вывод, что качество восстановления периода существенно зависит от структуры шума и при фиксированной структуре заметно не изменяется с ростом мощности алфавита.

Анализ данных, приведенных на рисунках *д* и *е*, позволяет проследить влияние уровня шума на качество распознавания в алфавитах различной мощности при равномерной структуре шума, когда все компоненты имеют одинаковый уровень. Рисунок *д* показывает, что для алфавитов первой группы даже при суммарном уровне шума в 10.5 % доля точно распознанных случаев не ниже 82.5 %, для алфавитов второй группы результаты хуже — шум в 10.5 % приводит только к качеству порядка 50 %.

Отметим также результаты для алфавита мощности 10 в первой и второй группах. Поскольку переход к мощности 10 осуществлялся в первой группе по фрагменту длины 16 в алфавите мощности 16, а во второй группе — по фрагменту длины 60 в алфавите мощности 60, то длина периодически повторяющегося фрагмента в последнем случае была приблизительно в полтора раза больше. Мы наблюдаем снижение качества распознавания во втором случае (см. рисунки *д* и *е*). Если для первой группы при равномерном шуме качество распознавания (доля зашумленных последовательностей с точно распознанным периодом) при суммарном уровне шума от 3 до 10.5 % составила от 100 до 92 %, то для той же мощности 10 во второй группе — уже от 85 до 58 %.

В целом, если предположить, что реальные данные подвержены шумам различных типов с близкой интенсивностью, и считать удовлетворительным 65 %-ное качество точного восстановления периода в смысле $\varepsilon_n(\bar{q}_n^\sigma, q_n^\sigma)$, то на основании рисунка и табл. 1 и 2 предложенный в [5] метод применим для алфавитов первой группы при уровне шума 10.5 % и для алфавитов второй группы — до уровня шума в 7.5 %.

Развитие данного исследования авторы видят в проведении дополнительного анализа зависимости квантилей $\varepsilon_n(\bar{q}_n^\sigma, q_n^\sigma)$ от параметров эксперимента.

Благодарности. Работа выполнена при финансовой поддержке РФФИ (гранты № 19-07-00150 и 19-07-00151).

Список литературы

[1] **Zhukova G., Smetanin Yu., Uljanov M.** Informative symbolic representations as a way to qualitatively analyses time series. Proceedings of the 2019 International Conference on Engineering Technologies and Computer Science: Innovation & Application. 2019: 43–47.

[2] **Lin J. et al.** A symbolic representation of time series, with implications for streaming algorithms. Proceedings of the 8th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery. ACM; 2003: 2–11.

[3] **Нестеренко А.Ю.** Алгоритмы поиска длин циклов в последовательностях и их приложения. Фундаментальная и прикладная математика. 2010; 16(6):109–122.

[4] **Скляр А.Я.** Анализ и устранение шумовой компоненты во временных рядах с переменным шагом. Кибернетика и программирование. 2019; (1):51–59. DOI:10.25136/2306-4196.2019.1.27031.

[5] **Жукова Г.Н., Жуков А.В., Сметанин Ю.Г., Ульянов М.В.** Метод определения периода зашумленной периодической символьной последовательности, основанный на позициях подслов в последовательности. Современные информационные технологии и ИТ-образование. 2020; 16(1):23–32. DOI:10.25559/SITITO.16.202001.23-32.

[6] **Ульянов М.В.** Подход к идентификации длины цикла в символьных последовательностях с шумом, основанный на энтропии слов. Современные технологии в науке и образовании: Сборник трудов III Международного науч.-техн. форума: в 10 т. Т. 4 / под общ. ред. О.В. Миловзорова. Рязань: Рязан. гос. радиотехн. ун-т; 2020: 124–128. ISBN:978-5-7722-0301-9.

[7] **Жукова Г.Н., Сметанин Ю.Г., Ульянов М.В.** Вероятностная модель шумов для периодических символьных последовательностей. Современные информационные технологии и ИТ-образование. 2019; 15(2):431–440. DOI:10.25559/SITITO.15.201902.431-440.

[8] **Левенштейн В.И.** Двоичные коды с исправлением выпадений, вставок и замещений символов. Доклады АН СССР. 1965; 163(4):845–848.

Вычислительные технологии, 2021, том 26, № 5, с. 95–105. © ФИЦ ИВТ, 2021 ISSN 1560-7534
Computational Technologies, 2021, vol. 26, no. 5, pp. 95–105. © FRC ICT, 2021 eISSN 2313-691X

INFORMATION TECHNOLOGIES

DOI:10.25743/ICT.2021.26.5.008

The influence of the cardinality of the alphabet on the quality of reconstruction of a symbolic periodic sequence from a sequence with noise

Zhukova Galina N.[1,*], Ulyanov Mikhail V.[2,3]

[1]HSE University, 109028, Moscow, Russia

[2]V.A. Trapeznikov Institute of Control Sciences of Russian Academy of Sciences, 117997, Moscow, Russia

[3]M.V. Lomonosov Moscow State University, 119991, Moscow, Russia

*Corresponding author: Zhukova Galina N., e-mail: gzhukova@hse.ru

Received August 6, 2021, revised August 12, 2021, accepted August 18, 2021.

Abstract

The relevance of this study is associated with the presence of a wide range of applied problems in real-world data processing and analysis. It is sensible to encode information using symbols from a finite alphabet in such problems. By varying the cardinality of the alphabet, in the description

of the process, the symbolic representation provides a level of detail sufficient for real-world data analysis. However, for a number of subject areas in which it is possible to use symbolic coding of trajectories of the examined processes researchers face the presence of distortions, noise, and fragmentation of information. This occurs in bioinformatics, medicine, digital economy, time series forecasting and analysis of business processes. Periodic processes are widely represented in these subject areas. Without noise, these processes correspond to periodic symbolic sequences, i. e. words over a finite alphabet. A researcher often receives a sequence distorted by noises of various origins as the experimental data, instead of the expected periodic symbolic sequence. Under these conditions, when solving the problem of identifying the periodicity, which includes both the determination of a periodically repeating symbolic fragment and its length, hereinafter called the period, the problem requires reducing the effect of noise on the experimental results.

The article deals with the problem of recovering periodic sequences, distorted by presence of noise along the replaced and deleted symbols. Since the level of detail in the description of the process depends on the cardinality of the alphabet, it is of interest to study the influence of the level of detail in the symbolic description on the possibility of recovering complete information about the initially periodic sequences.

The article experimentally examines the dependence of the cardinality of the alphabet on the quality characteristics of the period recovery method proposed by the authors. For alphabets of different cardinalities, the proportion of sequences with a satisfactorily reconstructed period and the relative error in determining the length of the period are given. The quality of reconstruction of a periodically repeating fragment is estimated by the ratio of the editing distance from the reconstructed periodic sequence to the original sequence distorted by noise.

Keywords: symbolic sequence, cardinality of an alphabet, periodic sequence, sequence with noise, noise of insertion, noise of deletion, noise of change.

Citation: Zhukova G.N., Ulyanov M.V. The influence of the cardinality of the alphabet on the quality of reconstruction of a symbolic periodic sequence from a sequence with noise. Computational Technologies. 2021; 26(5):95–105. DOI:10.25743/ICT.2021.26.5.008. (In Russ.)

Acknowledgements. This work is supported by RFBR (grants No. 19-07-00150 and 19-07-00151).

References

1. **Zhukova G., Smetanin Yu., Uljanov M.** Informative symbolic representations as a way to qualitatively analyses time series. Proceedings of the 2019 International Conference on Engineering Technologies and Computer Science: Innovation & Application. 2019: 43–47.
2. **Lin J. et al.** A symbolic representation of time series, with implications for streaming algorithms. Proceedings of the 8th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery. ACM; 2003: 2–11.
3. **Nesterenko A.Yu.** Cycle detection algorithms and their applications. Journal of Mathematical Sciences (New York). 2012; 182(4):518–526.
4. **Sklyar A.Ya.** Analysis and elimination of noise component in time series with variable step. Kibernetika i Programmirovanie. 2019: (1):51–59. DOI:10.25136/2306-4196.2019.1.27031.
5. **Zhukova G.N., Zhukov A.V., Smetanin Yu.G., Ulyanov M.V.** The method of estimating the period of a symbolic periodic sequence with noise, based on the sub-words positions in the sequence. Modern Information Technology and IT-education. 2020; 16(1):23–32. DOI:10.25559/SITITO.16.202001.23-32. (In Russ.)
6. **Ulyanov M.V.** Podkhod k identifikatsii dliny tsikla v simvol'nykh posledovatel'nostyakh s shumom, osnovannyy na entropii slov [An approach to identifying the cycle length in symbolic sequences with noise based on the entropy of words]. Ryazan: Ryazanskiy Gosudarstvennyy Radiotekhnicheskiy Universitet; 2020: 124–128. ISBN:978-5-7722-0301-9. (In Russ.)
7. **Zhukova G.N., Smetanin Yu.G., Ulyanov M.V.** A stochastic model of noises for periodic symbol sequences. Modern Information Technology and IT-education. 2019; 15(2):431–440. DOI:10.25559/SITITO.15.201902.431-440.
8. **Levenshtein V.I.** Binary codes capable of correcting deletions, insertions, and reversals. Dokl. Akad. Nauk SSSR. 1965; 163(4):845–848. (In Russ.)