ИНФОРМАТИКА

УДК 519.237.8

К. Ю. Староверова, В. М. Буре

МЕРА РАЗЛИЧИЯ ВРЕМЕННЫХ РЯДОВ, ОСНОВАННАЯ НА ИХ ХАРАКТЕРИСТИКАХ

Санкт-Петербургский государственный университет, Российская Федерация,
199034, Санкт-Петербург, Университетская наб., 7–9

Кластеризация временных рядов — актуальная задача современного анализа данных. Она требует разработки мер различия, учитывающих зависимость объектов от времени. В этой области уже была проделана большая работа, многие методы для нахождения расстояния между временными рядами основаны на сокращении размерности. Однако слабым местом на данный момент остается кластеризация коротких временных рядов, которые широко используются в таких областях как экономика, социология, демография и др. Поэтому предлагается новый метод, который строит матрицу расстояний только по характеристикам ряда. В статье описаны современные меры различия временных рядов; предложенная нами мера различия, основанная на характеристиках, и показаны результаты экспериментов применения этого метода на двух наборах искусственных данных в сравнении с другими методами. Библиогр. 12 назв. Ил. 2. Табл. 1.

Ключевые слова: кластеризация, мера различия временных рядов, классификация.

K. Yu. Staroverova, V. M. Bure

CHARACTERISTICS BASED DISSIMILARITY MEASURE FOR TIME SERIES

St. Petersburg State University, 7–9, Universitetskaya nab.,
St. Petersburg, 199034, Russian Federation

It is necessary to invent dissimilarity measures which take into account the temporal nature of a time series. Such measures can be utilized for classification and clustering of time series. Great work has been conducted on this problem, but most measures use dimensionality reduction techniques. Such methods give accurate results for big data, but demonstrate a weakness now in short time series clustering. Many fields such as economics, demography, sociology, and others are presented by short time series. That is why a new method based on time series characteristics is introduced here. It is based on time series characteristics which are split into three groups: constant, dynamic and behavioural. A researcher can control the influence of the characteristics of each group as a result. Besides, we present a brief description of up-to-date dissimilarity measures from the R environment. The results of experiments on two synthetic data sets and comparison of our measure and other up-to-date methods are then presented. Refs 12. Figs 2. Table 1.

Keywords: clustering, time series similarity measure, classification.

Староверова Ксения Юрьевна — студент; ksenygnirps@gmail.com
Буре Владимир Мансурович — доктор технических наук, профессор; vlb310154@gmail.com

Staroverova Kseniya Yurievna — student; ksenygnirps@gmail.com
Bure Vladimir Mansurovich — doctor of technical sciences, professor; vlb310154@gmail.com

DOI: 10.21638/11701/spbu10.2017.105

Введение. В современном мире для решения многих задач используются методы машинного обучения. К таким задачам относятся кластеризация и классификация данных. Большую роль при проведении кластеризации и классификации объектов играет выбор меры сходства или различия объектов для построения матрицы расстояний, так как неточно подобранная мера повлечет за собой неточные результаты решения задачи. Несмотря на то, что область анализа данных и машинного обучения развивается довольно стремительно, мы столкнулись с трудностью при исследовании неоднородности показателя заболеваемости в Санкт-Петерурге за последние 16 лет. Основная проблема состояла в подборе меры различия временных рядов, которую можно было бы использовать для кластеризации районов Санкт-Петербурга. Особенность, которую нужно учитывать при выборе меры различия временных рядов, состоит в том, что данные зависят от времени. Иногда постановка задачи позволяет эту особенность игнорировать и рассматривать временной ряд как вектор, однако чаще требуется сформировать группы, внутри которых объекты будут схожими именно в динамике.

В области кластеризации временных рядов была проделана большая работа: описано применение алгоритма динамической трансформации шкалы в кластерном анализе [1, 2]; этот же алгоритм, но с помощью теории скрытых марковских моделей предложен в [3]; использование коэффициентов автокорреляции, спектральных характеристик, вейвлет-коэффициентов отмечено в работах [4, 5], более полный обзор существующих методов представлен в статье [6].

Существующие меры различия дают хорошие результаты при работе с длинными рядами и зачастую даже применяют различные приемы по сокращению размерности, но при кластеризации коротких временных рядов такие меры могут давать недостаточно точные оценки близости объектов, что приводит к увеличению количества ошибок кластеризации. Нами предложена мера различия временных рядов — «мера различия, основанная на характеристиках» (MPOX), или “characteristics based dissimilarity measure” (CBDM), которая учитывает различие между статистическими характеристиками объектов кластеризации или классификации. Эксперименты на искусственных наборах данных показали, что выбор MPOX для кластеризации временных рядов оправдан, так как количество ошибок кластеризации оказалось меньше, чем при использовании многих других наиболее известных мер различия, время работы алгоритма оказалось также меньше. Построение матрицы MPOX-расстояний реализовано на языке R, в этом же пакете проведен сравнительный анализ данного метода с другими с помощью библиотек “cluster” и “TSclust”.

Обзор методов библиотеки “TSclust” пакета R. Библиотека представляет широкий выбор мер различия для кластерного анализа временных рядов и несколько наборов искусственных и реальных данных [7], поэтому она выбрана для проведения экспериментов и сравнительного анализа существующих мер с MPOX. Построение матрицы расстояний производится при помощи функции $diss(SERIES, METHOD, \dots)$, где $SERIES$ — это матрица, список или датафрейм с набором временных рядов, а $METHOD$ — строка с названием меры различия.

Существует условное разделение методов на 3 группы.

Свободные от предположений о модели методы строят расстояние непосредственно между наблюдениями двух рядов или некоторыми их характеристиками.

EUCL: евклидово расстояние, широко применяется при кластеризации, но не всегда может быть использовано, если объектом является временной ряд, в силу того, что зависимость наблюдений от времени игнорируется.

1. *ACF*: расстояние между коэффициентами автокорреляции рядов.
2. *PACF*: расстояние между частичными коэффициентами автокоррелиции рядов.
3. *COR*: вычисляет корреляцию Пирсона между наблюдениями объектов.
4. *CORT*: мера, учитывающая различие между наблюдениями, которое может быть представлено любым традиционным методом (расстояние Евклида, Минковского, Фреше, Манхеттенское), и разное поведение рядов.
5. *FRECHET*: расстояние Фреше, учитывает зависимость данных от времени, однако не принимает во внимание поведение временных рядов или характер их изменчивости.
6. *DTWARP*: алгоритм динамической трансформации шкалы, широко применяется при классификации и кластеризации речи, так как инвариантен относительно масштабирования и сдвига по временной оси, имеет тот же недостаток, что и предыдущий метод, кроме того, использование этого метода на любых данных неоправдано, так как иногда реальные различия в поведении могут быть не учтены или значительно уменьшены после преобразований временной оси.
7. *PER*: расстояние между периодограммами.
8. *INT.PER*: перед вычислением расстояния имеет преимущества интегрирование периодограмм. Это расстояние может быть нормировано, в таком случае большую роль играет форма кривой временного ряда.
9. *SPEC.GLK*: различие между двумя рядами оценивается в терминах значения статистики равенства логарифма спектров, для проверки гипотезы используется обобщенный критерий отношения правдоподобия [8].
10. *SPEC.ISD*: вычисляется как интеграл от квадратов разностей между непараметрическими оценками логарифма спектра.
11. *SPEC.LLR*: представлено отношением локальных линейных спектральных оценок.

Методы, основанные на предположениях о модели, строятся на приближении временных рядов моделями ARIMA.

1. *AR.MAH*: строится статистика, отражающая значимость различий процесов, для этого используются коэффициенты авторегрессии, дисперсия белого шума, матрицы ковариаций рядов. Статистика распределена как χ^2, нулевая гипотеза говорит о абсолютном сходстве процессов.
2. *AR.LPC.CEPS*: расстояние между ЛПК-кепстральными коэффициентами, т. е. кепстральными коэффициентами сигнала, которые вычисляются через коэффициенты авторегрессии.
3. *AR.PIC*: так как считается, что любой стационарный процесс можно аппроксимировать моделью $\mathrm{AR}(\infty)$, этот метод находит евклидово расстояние между коэффициентами модели, подобранной по критериям AIC или BIC.

Методы, основанные на оценке сложности временного ряда, отличаются от предыдущих тем, что они не учитывают временные характеристики и особенности процессов, не выдвигают предположения о модели, которой можно аппроксимировать процесс, зато оценивают уровень информации, содержащейся в каждом временном ряде.

1. *CDM*: мера, основанная на различии физического размера сжатых временных рядов. Допускается выбор одного из трех алгоритмов сжатия: “gzip”, “bzip2”,

"xz". Возможно использование символьного представления процесса, например SAX-методов.

2. *NCD*: отличается от предыдущего метода только формулой вычисления расстояния, авторы [9] называют данное расстояние «нормализованным».
3. *CID*: скорректированное евклидово расстояние. Корректировка происходит путем домножения на коэффициент, определяющий отношение сложностей двух рядов.
4. *MINDIST.SAX*: сначала происходит нормализация временного ряда, потом сокращение размерности алгоритмом PAA (кусочно-агрегирующая аппроксимация), после — преобразование алгоритмом SAX в символьный ряд и вычисление расстояния путем нахождения минимума.
5. *PDC*: сначала выполняется перестановка наблюдений в восходящем порядке, после чего строится код, отражающий распределение упорядоченных наблюдений на первоначальном ряде, этот код используется для оценки сложности процесса. Расстояние между двумя рядами вычисляется как различие между построенными кодами.

Важно отметить, что каждый метод обладает своей спецификой. Во-первых, выбор одной из трех групп должен быть не случайным и соответствовать целям исследования, во-вторых, выбор меры также должен быть обусловлен спецификой задачи. Например, если необходимо производить кластеризацию длинных рядов ежесекундно, при этом визуализировать данные, а точностью можно пренебречь, то стоит отдать предпочтение более простым методам, таким как вычисление евклидова расстояния или коэффициентов корреляции Пирсона, и наоборот, если проводится исследование, в котором важна точность результатов, то выбирать нужно более сложные методы, например динамическое преобразование шкалы, методы, основанные на спектре процесса, преобразованиях вейвлетами.

Описание метода MPOX. Создание этого метода началось с формулировки недостатка существующих методов: многие из них не предназначены для работы с короткими рядами. Вопрос кластеризации временных рядов малой размерности является актуальным, так как статистика по многим экономическим, социальным, демографическим и другим показателям собирается нечасто — один раз в год, квартал или месяц и доступна за небольшие периоды в несколько лет. Именно поэтому перед нами стоял вопрос: каким образом можно представить информацию о временном ряде, кроме уже существующих способов, так, чтобы она являлась достаточно точным описанием и для коротких временных рядов.

Метод основан на вычислении характеристик временного ряда, которые делятся на три группы: не зависящие от времени, описывающие динамику и определяющие изменчивость. Для каждой рассчитывается величина, отражающая различие временных рядов по характеристикам из конкретной группы, а затем строится линейная их комбинация.

Перед вычислением характеристик, *не зависящих от времени*, преобразуем временные ряды, сохраняя различие между объектами в плоскости, но отображая значения наблюдений в отрезок $[0, 1]$, т. е., если $M = [n \times m]$ — это матрица n временных рядов длиной m, тогда преобразование определяется формулой

$$\overline{M} = \frac{M - \min M}{M - \max M}.$$

После этого по матрице $\overline{M}$ для каждого временного ряда (строки) вычисляем

матрицу характеристик $C = [n \times 5]$: столбцы $C_{*,1}$ — средних значений каждого ряда, $C_{*,2}$ — стандартных отклонений, $C_{*,3}$ — медиан, $C_{*,4}$ — минимумов, $C_{*,5}$ — максимумов. Здесь не учитывается зависимость данных от времени.

Для вычисления *различий в динамике* для каждого временного ряда строим четыре вектора, характеризующих его поведение. Перед этим выполним преобразование матрицы M, но в этот раз таким образом, чтобы максимальному значению каждого ряда соответствовала 1, а минимальному — 0:

$$\widetilde{M}_{k,*} = \frac{M_{k,*} - \min M_{k,*}}{\max M_{k,*} - \min M_{k,*}}, \tag{1}$$

где $M_{k,*}$ означает строку матрицы M с номером k. Определим первые разности ряда с лагом l

$$\widetilde{FD}_{k,t}(l) = \widetilde{M}_{k,t+l} - \widetilde{M}_{k,t}, \tag{2}$$

где $k = \overline{1,n}$; $t = \overline{1,m-l}$. Для описания динамики временного ряда будем использовать первые разности с лагом 1

$$\widetilde{FD}_{k,t} = \widetilde{FD}_{k,t}(1).$$

Нормализация (1) позволяет не обращать внимание на разбросанность наблюдений в пространстве и исследовать только поведение ряда. Построим первую матрицу, каждая строка которой отражает динамику процесса, а именно, показывает участки неубывания

$$D^1_{k,t} = \begin{cases} 1, \text{ если } \widetilde{FD}_{k,t} \geqslant 0, \\ 0, \text{ иначе,} \end{cases}$$

где $k = \overline{1,n}$; $t = \overline{1,m-1}$. Вторая матрица показывает, как временной ряд флуктуирует около среднего значения:

$$D^2_{k,t} = \begin{cases} 1, \text{ если } \widetilde{M}_{k,t} \geqslant \mathrm{E}(\widetilde{M}_{k,*}), \\ 0, \text{ иначе,} \end{cases}$$

где $k = \overline{1,n}$; $t = \overline{1,m}$. Третья и четвертая матрицы показывают бóльшие отклонения от среднего значения, чем стандартное отклонение:

$$D^3_{k,t} = \begin{cases} 1, \text{ если } \widetilde{M}_{k,t} \geqslant \mathrm{E}(\widetilde{M}_{k,*}) + \mathrm{Var}(\widetilde{M}_{k,*}), \\ 0, \text{ иначе,} \end{cases}$$

$$D^4_{k,t} = \begin{cases} 1, \text{ если } \widetilde{M}_{k,t} \leqslant \mathrm{E}(\widetilde{M}_{k,*}) - \mathrm{Var}(\widetilde{M}_{k,*}), \\ 0, \text{ иначе,} \end{cases}$$

где $k = \overline{1,n}$; $t = \overline{1,m}$; $\mathrm{E}(\widetilde{M}_{k,*})$ — среднее значение временного ряда k; $\mathrm{Var}(\widetilde{M}_{k,*})$ — его дисперсия.

Характеристики, отвечающие за *изменчивость*, включают в себя 15 величин, которые вычисляются для первых разностей ряда с лагом 1 и 2, их можно представить в виде матрицы $V = [n \times 15]$. Определим среднюю скорость роста $V_{*,1}$, спада $V_{*,2}$ и изменения ряда $V_{*,3}$. Пусть $l = 1$, если количество участков возрастания и убывания ряда

$$m_k^g(l) = \sum_{t=1}^{m-l} \mathrm{I}\{\widetilde{FD}_{k,t}(l) > 0\}, \tag{3}$$

$$m_k^d(l) = \sum_{t=1}^{m-l} \mathrm{I}\{\widetilde{FD}_{k,t}(l) < 0\}, \tag{4}$$

где I — это индикатор (принимает значение 1, когда верно выражение в скобках, иначе — 0), тогда компоненты столбцов $V_{*,1},\ V_{*,2},\ V_{*,3}$ задаются как

$$V_{k,1}(l) = \frac{1}{m_k^g(l)} \sum_{t=1}^{m-l} \widetilde{FD}_{k,t}(l)\, \mathrm{I}\{\widetilde{FD}_{k,t}(l) > 0\}, \tag{5}$$

$$V_{k,2}(l) = \frac{1}{m_k^d(l)} \sum_{t=1}^{m-l} \widetilde{FD}_{k,t}(l)\, \mathrm{I}\{\widetilde{FD}_{k,t}(l) < 0\}, \tag{6}$$

$$V_{k,3} = \mathrm{E}(\widetilde{FD}_{k,*}(l)), \tag{7}$$

где $k = \overline{1,n}$. Вычислим первые разности с лагом $l = 2$ по формуле (2), тогда компоненты столбцов $V_{*,4},\ V_{*,5},\ V_{*,6}$ можно вычислить для $\widetilde{FD}_{k,t}(2)$ по формулам (3)–(7).

Следующие характеристики показывают величину наибольшего роста и наибольшего спада, процесс их вычисления приведен в виде алгоритма:

Вход: FD — матрица первых разностей, где строка с номером k соответствует первой разности временного ряда k;
Выход: для каждого ряда с номером k пара значений $\{Growth[\,k\,], Decline[\,k\,]\}$;

$n \leftarrow$ количество строк FD
$m \leftarrow$ количество столбцов FD
Для i **от** 1 **до** n
    $Growth[\,i\,] \leftarrow 0$
    $Decline[\,i\,] \leftarrow 0$
    $Sum \leftarrow FD[\,i,\ 1\,]$
    Для j **от** 2 **до** m
        Если $(FD[\,i,\ j\,] \cdot FD[\,i-1,\ j\,]) < 0$ **И** $Sum > Growth[\,i\,]$
            То $Growth[\,i\,] \leftarrow Sum$
                $Sum \leftarrow FD[\,i,\ j\,]$
        Если $(FD[\,i,\ j\,] \cdot FD[\,i-1,\ j\,]) < 0$ **И** $Sum < Decline[\,i\,]$
            То $Decline[\,i\,] \leftarrow Sum$
                $Sum \leftarrow FD[\,i,\ j\,]$
        Если $(FD[\,i,\ j\,] \cdot FD[\,i-1,\ j\,]) > 0$
            То $Sum[\,i\,] \leftarrow Sum + FD[\,i,\ j\,]$

Столбцы $V_{*,7},\ V_{*,8}$ рассчитываются по описанному выше алгоритму для первых разностей ряда с лагом 1, а столбцы $V_{*,9},\ V_{*,10}$ — для первых разностей ряда с лагом 2.

Последние 5 столбцов матрицы V соответствуют минимальному значению, квартилям и максимальному значению первых разностей с лагом 1.

Перед тем как определить расстояние между временными рядами, введем функцию, которая оценивает расстояние между характеристиками динамики. Пусть M_k — это строка матрицы M с номером k, т. е. временной ряд, тогда

$$\mathrm{DIST}_{Dyn}(M_{k_1}, M_{k_2}) = \frac{1}{4}\left(\frac{1}{m-1}\sum_{t=1}^{m-1} D^i_{k_1,t} \oplus D^i_{k_2,t} + \sum_{i=2}^{4}\frac{1}{m}\sum_{t=1}^{m} D^i_{k_1,t} \oplus D^i_{k_2,t}\right),$$

где $D^i_{k_1,t} \oplus D^i_{k_2,t}$ — покомпонентное сложение по модулю 2. Расстояние между временными рядами складывается из трех величин: различие между характеристиками, не зависящими от времени, динамики и изменчивости:

$$\begin{aligned}\mathrm{DIST}_{CBM}(M_{k_1}, M_{k_2}) = \alpha\sqrt{\sum_{i=1}^{5}(C_{k_1,i} - C_{k_2,i})^2} + \beta\,\mathrm{DIST}_{Dyn}(M_{k_1}, M_{k_2}) + \\ + (1-\alpha-\beta)\sqrt{\sum_{i=1}^{15}(V_{k_1,i} - V_{k_2,i})^2}.\end{aligned} \tag{8}$$

Влияние каждой величины регулируется с помощью коэффициентов α и β, причем $(\alpha+\beta) \in [0,1]$. Расстояние, введенное в формуле (8), является метрикой.

Эксперимент. Идея использовать характеристики временных рядов для проведения кластеризации уже была описана в [10], однако удовлетворительных результатов получено не было. Нами проведен эксперимент на том же наборе искусственных данных, который содержит 600 временных рядов, сгенерированных по следующим моделям $y(t)$, $t = \overline{1,60}$:

1) нормальная модель: $y(t) = m + rs$, где $m = 30, s = 2, r \in [-3,3]$;
2) циклическая модель: $y(t) = m + rs + a\sin\frac{2\pi t}{T}$, где $a, T \in [10,15]$;
3) модель с восходящим трендом: $y(t) = m + rs + gt$, где $g \in [0.2, 0.5]$;
4) модель с нисходящим трендом: $y(t) = m + rs - gt$, где $g \in [0.2, 0.5]$;
5) модель со сдвигом вверх: $y(t) = m + rs + kx$, где $x \in [7.5, 20], k = 0$ до момента $\hat{t}$ и $k = 1$ после этого момента, $\hat{t} \in [20, 40]$;
6) модель со сдвигом вниз: $y(t) = m + rs - kx$, где $x \in [7.5, 20], k = 0$ до момента $\hat{t}$ и $k = 1$ после этого момента, $\hat{t} \in [20, 40]$.

Примеры временных рядов каждой модели представлены на рис. 1, *а*—*е*. Сравнивать метод MPOX с другими расстояниями можно по проценту правильно кластеризованных объектов P. В [10] этот показатель составил 47.3%. При кластеризации методом MPOX с коэффициентами $\alpha = 0.55$, $\beta = 0.1$ эффективность кластеризации $P = 85.2\%$. На рис. 2, *а* показано, как должны быть распределены объекты по кластерам, а на рис. 2, *б* — результаты, полученные рассматриваемым методом.

Сравнение MPOX с методами пакета "TSclust" показало, что для этого набора данных предлагаемый нами метод дает лучшие результаты. Для большей части методов $P < 60\%$ и только для алгоритма динамической трансформации шкалы (ДТШ) $P = 84.7\%$, однако время построения матрицы расстояний методом MPOX составляет 16.04 с, а для ДТШ — 445.68 с. Распределение объектов по кластерам показано на рис. 2, *в*.

Также можно сравнить индекс оценки силуэта кластеров, который часто применяется для оценки качества кластеризации и в нашем случае может говорить о

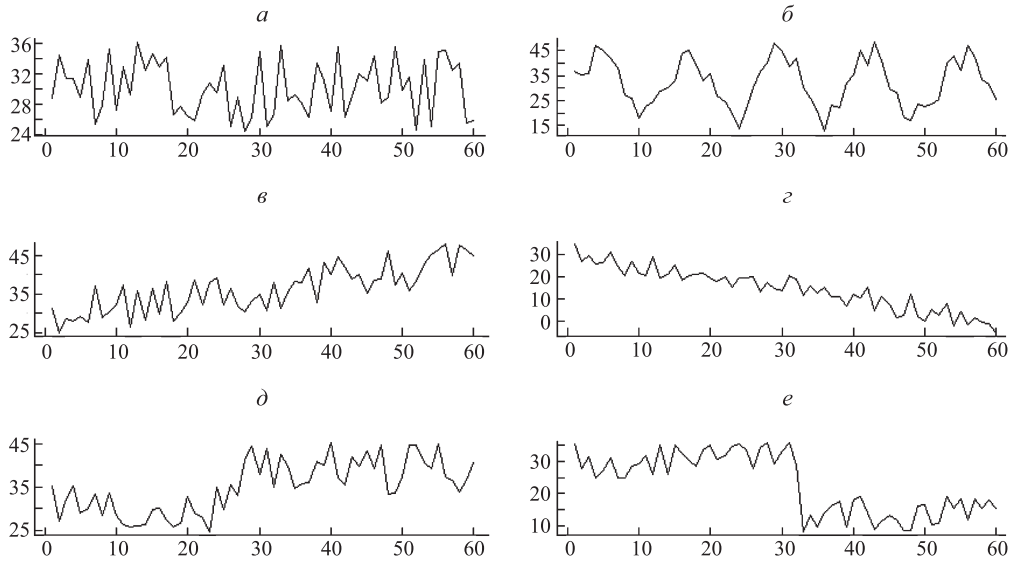

Рис. 1. Пример реализаций разных моделей

По оси абсцисс — время, по оси ординат — значение временного ряда.
Модели: *а* — нормальная, *б* — циклическая, *в* — с восходящим трендом, *г* — с нисходящим трендом, *д* — со сдвигом вверх, *е* — со сдвигом вниз.

некоторой устойчивости, так как высокий индекс оценки силуэта свидетельствует о том, что расстояние между объектами внутри кластера мало, а между элементами соседних кластеров — велико [11]. Индекс оценки силуэта может изменяться от 0 до 1, чем ближе значение к 1, тем лучше произведена кластеризация. Как следует из таблицы, в среднем индекс выше при кластеризации методом MPOX.

Индекс оценки силуэтов для кластеризаций методами MPOX и ДТШ

Метод	Кластер 1	Кластер 2	Кластер 3	Кластер 4	Кластер 5	Кластер 6
MPOX	0.39	0.56	0.13	0.42	0.50	0.13
ДТШ	0.58	0.22	0.14	0.12	0.28	0.25

Второй эксперимент был проведен с искусственными данными библиотеки “TSclust”. Набор содержит по 3 реализации длиной 200 каждой из таких моделей:

1) $X_t = 0.6X_{t-1} + \epsilon_t$;
2) $X_t = (0.3 - 0.2\epsilon_{t-1})X_{t-1} + 1 + \epsilon_t$;
3) $X_t = (0.9\exp(-X_{t-1}^2) - 0.6)X_{t-1} + 1 + \epsilon_t$;
4) $X_t = (0.3X_{t-1} + 1)I(X_{t-1} \geqslant 0.2) - (0.3X_{t-1} - 1)I(X_{t-1} < 0.2) + \epsilon_t$;
5) $X_t = 0.7|X_{t-1}|(2 + |X_{t-1}|)^{-1} + \epsilon_t$;
6) $X_t = 0.8X_{t-1} - 0.8X_{t-1}(1 + \exp(-10X_{t-1}))^{-1} + \epsilon_t$.

Эксперименты на этих данных уже проводились в [12], наилучшие результаты показали методы “INT.PER” $P = 88.9\%$ и “SPEC.LLR” $P = 83.3\%$. Примененный нами метод показывает результаты немного хуже, чем предыдущие два, однако лучше, чем остальные: $P = 82.8\%$ при $\alpha = 0.3,\ \beta = 0.3$.

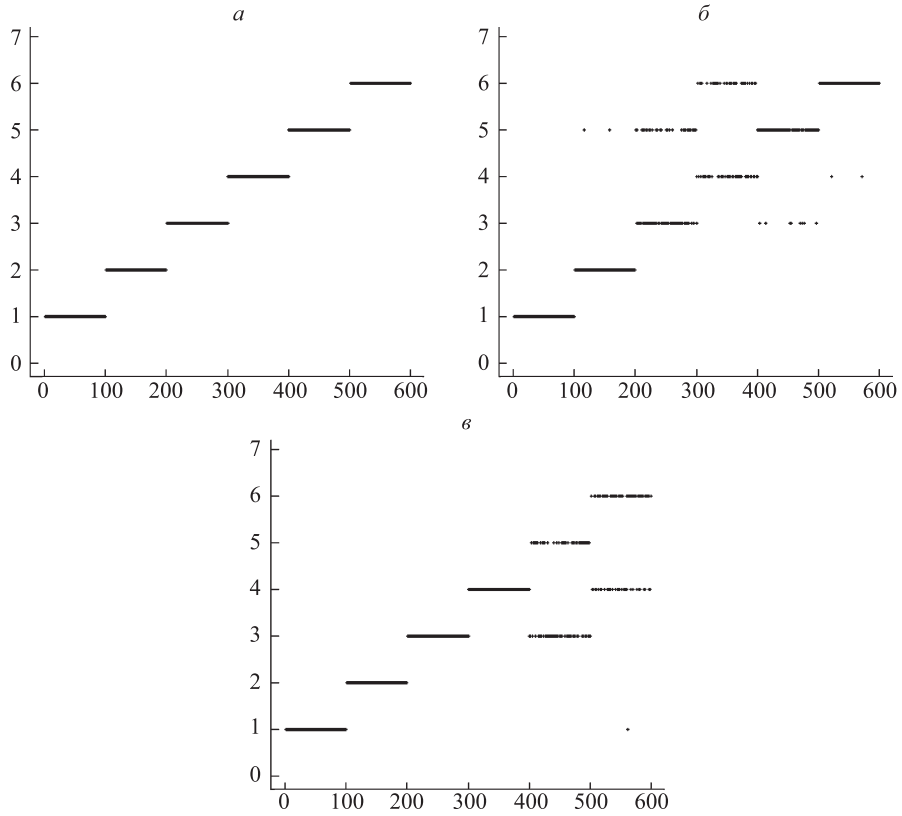

Рис. 2. Распределение объектов по кластерам

По оси абсцисс — номер объекта, по оси ординат — номер кластера.
Кластеризация: *а* — теоретическая, *б* — методом МРОХ, *в* — методом ДТШ.

Заключение. Метод, основанный на вычислении характеристик, показал неплохие результаты при проведении экспериментов, что позволяет использовать его в задачах кластерного анализа временных рядов. Преимуществами являются простота и скорость работы алгоритма. Стоит также отметить, что данный метод учитывает и динамику временного ряда и характер изменчивости, при этом предложенная мера различия является метрикой. Сейчас недостаток алгоритма заключается в подборе параметров α и β, поэтому естественным развитием становится разработка критериев по выбору этих коэффициентов. Однако постановка некоторых задач позволяет понять, какие характеристики оказываются более важными для кластеризации, в таком случае трудности с подбором коэффициентов не возникают.

Литература

1. *Sankoff D., Kruskal J.* Time warps, string edits, and macromolecules: the theory and practice of sequence comparison. Ontario: Addison Wesley Publ. Company, 1983. 382 p.

2. *Berndt D. J., Clifford J.* Using dynamic time warping to find patterns in time series // KDD workshop on knowledge discovery in databases. 1994. P. 359–370.

3. *Oates T., Firoiu L., Cohen P. R.* Clustering time series with hidden Markov models and dynamic time warping // Proc. of the IJCAI-99 workshop on neural, symbolic and reinforcement learning methods for sequence learning. 1999. P. 17–21.

4. *Maharaj E. A.* A significance test for classifying arma models // Journal of Statistical Computation and Simulation. 1996. Vol. 54, N 4. P. 305–331.

5. *Corduas M., Piccolo D.* Time series clustering and classification by the autoregressive metric // Computational Statistics & Data Analysis. 2008. Vol. 52, N 4. P. 1860–1872.

6. *Fu T. C.* A Review on time series data mining // Engineering Applications of Artificial Intelligence. 2011. Vol. 24, N 1. P. 164–181.

7. *Montero P. M., Vilar J. A.* Time series clustering utilities. Feb. 2015. URL: https://cran.r-project.org/web/packages/TSclust/TSclust.pdf (дата обращения: 01.11.2016).

8. *Fan J., Zhang W.* Generalised likelihood ratio tests for spectral density // Biometrika. 2004. Vol. 91, N 1. P. 195–209.

9. *Cilibrasi R., Vitànyi P. M.* Clustering by compression // IEEE Transactions on Information Theory. 2005. Vol. 51, N 4. P. 1523–1545.

10. *Alcock R. J., Manolopoulos Y.* Time-series similarity queries employing a feature-based approach // 7th Hellenic Conference on Informatics. 1999. P. 27–29.

11. *Сивоголовко Е. В.* Методы оценки качества четкой кластеризации // Компьютерные инструменты в образовании. 2011. № 4. С. 14–31.

12. *Montero P., Vilar J.* TSclust: An R package for time series clustering // Journal of Statistical Software. 2015. N 62.1. P. 1–43.

Для цитирования: Староверова К. Ю., Буре В. М. Мера различия временных рядов, основанная на их характеристиках // Вестник Санкт-Петербургского университета. Прикладная математика. Информатика. Процессы управления. 2017. Т. 13. Вып. 1. С. 51–60. DOI: 10.21638/11701/spbu10.2017.105

References

1. Sankoff D., Kruskal J. *Time warps, string edits, and macromolecules: the theory and practice of sequence comparison.* Ontario, Addison Wesley Publ. Company, 1983, 382 p.

2. Berndt D. J., Clifford J. Using dynamic time warping to find patterns in time series. *KDD workshop on knowledge discovery in databases*, 1994, pp. 359–370.

3. Oates T., Firoiu L., Cohen P. R. Clustering time series with hidden Markov models and dynamic time warping. *Processing of the IJCAI-99 workshop on neural, symbolic and reinforcement learning methods for sequence learning,* 1999, pp. 17–21.

4. Maharaj E. A. A significance test for classifying arma models. *Journal of Statistical Computation and Simulation,* 1996, vol. 54, no. 4, pp. 305–331.

5. Corduas M., Piccolo D. Time series clustering and classification by the autoregressive metric. *Computational Statistics & Data Analysis,* 2008, vol. 52, no. 4, pp. 1860–1872.

6. Fu T. C. A Review on time series data mining. *Engineering Applications of Artificial Intelligence,* 2011, vol. 24, no. 1, pp. 164–181.

7. Montero P. M., Vilar J. A. Time series clustering utilities. Feb. 2015. Available at: https://cran.r-project.org/web/packages/TSclust/TSclust.pdf (accessed 01.11.2016).

8. Fan J., Zhang W. Generalised likelihood ratio tests for spectral density. *Biometrika,* 2004, vol. 91, no. 1, pp. 195–209.

9. Cilibrasi R., Vitànyi P. M. Clustering by compression. *IEEE Transactions on Information Theory*, 2005, vol. 51, no. 4, pp. 1523–1545.

10. Alcock R. J., Manolopoulos Y. Time-series similarity queries employing a feature-based approach. *7th Hellenic Conference on Informatics*, 1999, pp. 27–29.

11. Sivogolovko E. V. Metodi ocenki kachestva chetkoi klasterizacii [Methods of estimation of smooth clustering quality]. *Komp'yuternye instrumenty v obrazovanii* [*Computer tools in education*], 2011, issue 4, pp. 14–31. (In Russian)

12. Montero P., Vilar J. TSclust: An R package for time series clustering. *Journal of Statistical Software,* 2015, no. 62.1, pp. 1–43.

For citation: Staroverova K. Yu., Bure V. M. Characteristics based dissimilarity measure for time series. *Vestnik of Saint Petersburg University. Applied mathematics. Computer science. Control processes*, 2017, volume 13, issue 1, pp. 51–60. DOI: 10.21638/11701/spbu10.2017.105

Статья рекомендована к печати проф. Л. А. Петросяном.

Статья поступила в редакцию 3 ноября 2016 г.

Статья принята к печати 19 января 2017 г.