ИСЛЕДОВАНИИЕ ВРЕМЕННЫХ РЯДОВ*

А. В. Грабовой[1], **В. В. Стрижов**[2]

Аннотация: Работа посвящена поиску периодических сигналов во временном ряду. Предлагается метод кластеризации точек временного ряда для поиска характерных переодических сигналов внутри временного ряда. Для построения признакового описания используется локальное снижения размерности фазового пространства при помощи метода главных компонент. Для оценки близости двух периодических сигналов рассматриивается расстояние между базисными векторами полученых методом главных компонет. Используя матрицу попарных растоний между точками временного ряда выполняется кластеризация точек временного ряда. Для анализа качества представленного алгоритма проводятся экпериенты на синтетических данных.

Ключевые слова: временные ряды; кластеризация временных рядов

DOI: 00.00000/00000000000000

1 Введение

Решается задача поиска локальных периодических сигналов акселерометра.

---

*Работа выполнена при поддержке РФФИ и правительства РФ.

[1]Московский физико-технический институт, grabovoy.av@phystech.edu

[2]Вычислительный центр им. А. А. Дородницына ФИЦ ИУ РАН, strijov@ccas.ru

Временные ряды это объекты сложной структуры, при классиифика-циии которых большую роль играет построение признакового пространства. Для этой цели возможно использование экспертного задания базовых функций и метода построения признаков на основе гипотезы порождения данных. В работе [1] рассматривается комбинированное признаковое описание на основе этих двух методов. В статье [2] рассматривается построеиние признаков и предлагается критерий избыточности выбранных признаков.

Данная работа посвящена поиску класификации сигналов внутри временного ряда. Предлагается локально использовать метод главных компонент для выделения базисных векторов, которые апроксимируют данный участок временного ряда и рассматриивается как признаковое описание этого участва. Используя признаковое описание временного ряда производиться их кластеризация.

Для решения данной задачи вводится ряд предположений о временном ряде. Предполагается, что периоды всех различных сигналов различаются не значительно, причем известен максимальный период сигналов и количество различных сигналов внутри временного ряда. Также предполгается, что класс сигнала во времени меняется не очень часто, а также что фазовые траектори сигналов из разных классов являются различными.

Проверка и анализ метода проводится на синтетической выборке.

2 Постановка

Задан временной ряд:

$$\mathbf{X} \in \mathbb{R}^{\mathrm{N}\times 1}. \tag{2.1}$$

Пусть временной ряд состоит из последовательности сигналов из множества $\mathcal{V}$:

$$\mathbf{X} = [\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_M], \tag{2.2}$$

где $\mathbf{v}_i$ некоторый сигнал из множества возможных сигналов $\mathcal{V}$. Причем $\forall i$ выполняется или $\mathbf{v}_i = \mathbf{v}_{i-1}$ или $\mathbf{v}_i = \mathbf{v}_{i+1}$. Пусть множество $\mathcal{V}$ удовлетворяет следующим свойствам:

$$|\mathcal{V}| = K, \quad \forall \mathbf{v} \in \mathcal{V} \ \ |\mathbf{v}| \leq T, \tag{2.3}$$

где $|\mathcal{V}|$ мощность множества сигналов, а $|\mathbf{v}|$ длина сигнала.

Расмотрим отображение:

$$a : x \to \{1, \cdots, K\}, \tag{2.4}$$

где $x \in \mathbf{X}$ некоторая точка временного ряда.

Требуется, чтобы отображение удовлетворяло следующим свойствам:

$$\begin{cases} a(x_1) = a(x_2), & \text{если } \exists \mathbf{v} \in \mathcal{V} : x_1, x_2 \in v \\ a(x_1) \neq a(x_2), & \text{если } \nexists \mathbf{v} \in \mathcal{V} : x_1, x_2 \in v \end{cases}$$

3 Кластеризация точек

Рассмотрим фазовую траекторию ряда $\mathbf{X}$:

$$\mathcal{H} = \{\mathbf{h}_t | \mathbf{h}_t = [x_{t-T}, x_{t-T+1}, \cdots, x_t],\ \mathrm{T} \leq t \leq \mathrm{N}\}. \tag{3.1}$$

Фазовая траектория разбиваеся на фазовые подпространства из $2T$ векторов:

$$\mathcal{S} = \{\mathbf{s}_t | \mathbf{s}_t = [h_{t-2T}, h_{t-2T+1}, \cdots, h_t],\ 2\mathrm{T} \leq t \leq \mathrm{N}\}. \tag{3.1}$$

Каждое Т-мерное подпространство s_t спроектируем на плоскость при помощи метода главных компонент. Получим представление базисных векторов плоскости, а также собственые числа, которые соответсвуют данным базисным векторам каждого подпраcтранства s_t в Т-мерном пространстве:

$$\mathcal{W} = \{\mathbf{W}_t | \mathbf{W}_t = [\mathbf{w}_t^1, \mathbf{w}_t^2]\}, \quad \mathcal{L} = \{\boldsymbol{\lambda}_t | \boldsymbol{\lambda}_t = [\lambda_t^1, \lambda_t^2]\}, \tag{3.3}$$

где $[\mathbf{w}_t^1, \mathbf{w}_t^2]$ и $[\lambda_t^1, \lambda_t^2]$ это базисные векторы и соответствующие им собственные числа плоскости построенной при помощи метода главных компонент для подпространстве s_t.

Рассмотрим расстояниие между элементами $\mathcal{W}$:

$$\rho(\mathbf{W}_1, \mathbf{W}_2) = \max_{\{\mathbf{a}, \mathbf{b}, \mathbf{c}\} \subset \mathbf{W}_1 \cup \mathbf{W}_2} V(\mathbf{a}, \mathbf{b}, \mathbf{c}), \tag{3.4}$$

где $V(\mathbf{a}, \mathbf{b}, \mathbf{c})$ — объем паралелепипеда построенного на векторах $\mathbf{a}, \mathbf{b}, \mathbf{c}$.

$\rho(\mathbf{W}_1, \mathbf{W}_2)$ является псевдометрикой в пространстве $\mathcal{W}$. Также $\rho(\mathbf{W}_1, \mathbf{W}_2)$ является метрикой, если дополниительно указать, что базисы соответсвующие паралельным плоскостям не различимы.

Рассмотрим растояние между элементами $\mathcal{L}$:

$$\rho\left(\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2\right) = \sqrt{\left(\boldsymbol{\lambda}_1 - \boldsymbol{\lambda}_2\right)^{\mathsf{T}}\left(\boldsymbol{\lambda}_1 - \boldsymbol{\lambda}_2\right)}. \tag{3.5}$$

$\rho\left(\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2\right)$ является метрикой в пространстве $\mathcal{L}$.

Матрица попарных растояний между базисными векторами для временного ряда $\mathbf{X}$:

$$\mathbf{M}_c = [0, 1]^{N \times N}. \tag{3.6}$$

Матрица попарных растояний между собственными значениями для временного ряда $\mathbf{X}$:

$$\mathbf{M}_l = [0, 1]^{N \times N}. \tag{3.7}$$

Используя выражения (3.4-7) определим растояниие между двумя точками t_1, t_2 временного ряда:

$$\rho\left(t_1, t_2\right) = \rho\left(\mathbf{W}_1, \mathbf{W}_2\right) + \rho\left(\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2\right), \quad \mathbf{M} = \mathbf{M}_l + \mathbf{M}_c, \tag{3.8}$$

где $\rho\left(t_1, t_2\right)$ является метрикой, как сумма двух метрик. Матрица $\mathbf{M}$ является матрицей попарных растояний между двумя точками временного ряда.

Используя матрицу попарных растояний $\mathbf{M}$ выполним кластеризацию моментов времени временного ряда, получим следующее отображение:

$$a : x \to \{1, \cdots, K\}, \tag{3.9}$$

где x некоторая точка временного ряда $\mathbf{X}$.

4 Эксперимент

Для анализа свойст предложенного алгоритма был проведен вычислительный жксперимент в котором кластеризация точек временного ряда проводилась используя матрицы попарных растояний $(3.6-8)$.

В качестве данных использовались две выборки временных ряжов. Выборка "найти хорошую реальную выборку"это реальные временые ряды.

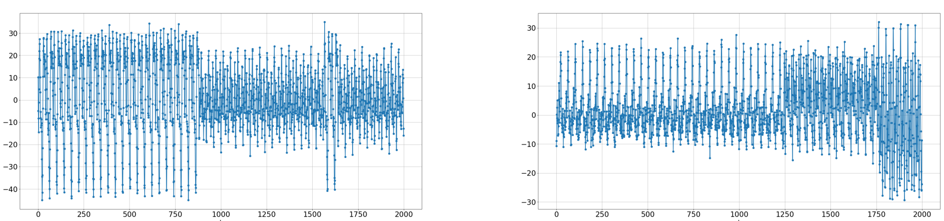
Синтетические временные ряды были построены при помощи обрещаного ряда Фурье с произвольными коэфициентами. Генерация данных состояла из двух этапов. На первом этапе генерировались короткие сигналы $\mathbf{v}$ для построения множества $\mathcal{V}$. Вторым этапом генерации выборки $\mathbf{X}$ является следующим случайным процесом:

$$\mathbf{X} = [\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_M], \quad \begin{cases} \mathbf{v}_1 \sim \mathcal{U}(\mathcal{V}), \\ \mathbf{v}_i = \mathbf{v}_{i-1}, & \text{с вероятностью } \frac{3}{4}, \\ \mathbf{v}_i \sim \mathcal{U}(\mathcal{V}), & \text{с вероятностью } \frac{1}{4} \end{cases} \tag{4.1}$$

где $\mathcal{U}(\mathcal{V})$ — равномерное распределение на объектах из $\mathcal{V}$.

Таблица 1: Описание выборок

Выборка	N	K	T
Real			
Synthetic 1	2000	2	20
Synthetic 2	2000	3	20

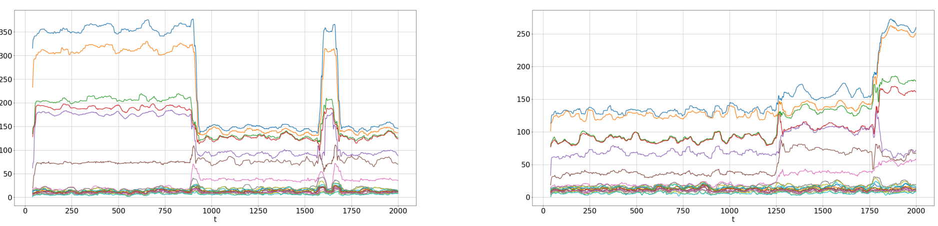


(a) Synthetic 1 (b) Synthetic 2

Рис. 1: Пример синтетически построеных временных рядов

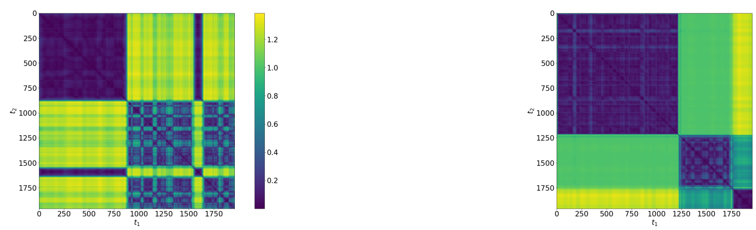
Синтетические данные. На рис. 1 приведен пример синтетически построенных временных рядов. На рис. 1a показан пример ряда в котором количество сигналов $K = 2$, а длина каждого сигнала $T = 20$. На рис. 1b показан пример ряда в котором количество сигналов $K = 3$, а длина каждого сигнала $T = 20$.

На рис. 2 показан график зависимости значения сингулярных чисел локальной апроксиимации с течением времени. Значение сингулярных



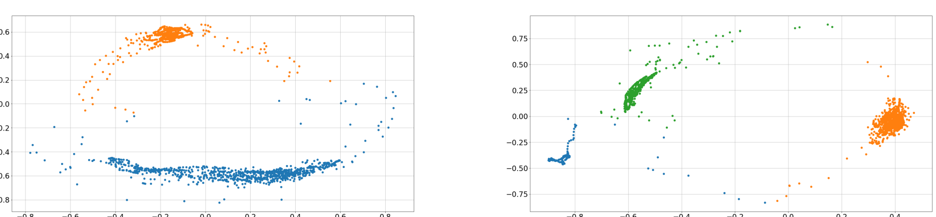
(a) Synthetic 1 (b) Synthetic 2

Рис. 2: График зависимостии значения сингулярных чисел метода главных компонент



(a) Synthetic 1 (b) Synthetic 2

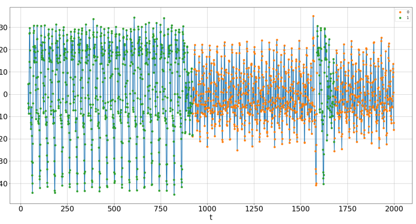
Рис. 3: Матрица попарных растояний $\mathbf{M}$ между точками временного ряда



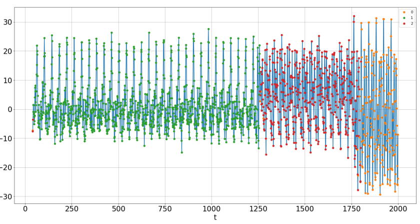
(a) Synthetic 1 (b) Synthetic 2

Рис. 4: Проекция точек временного на плоскость при помощи матрицы попарных растояний $\mathbf{M}$

чисел, которые соответствуют первым двум главным компонентам значительно меняются с течением времени t.



(a) Synthetic 1



(b) Synthetic 2

Рис. 5: Кластеризация точек временного ряда

На рис. 3 проилюстрированы матрицы попарных растояний **M** между построены при помощи формулы (3.8). Используя матрицу попарных растояний и метод Multidimensional Scaling [3] визуальзуализируем точки временого ряда на плоскости. На рис. 4 показана визуализация точек на плоскости и выполнена их кластеризация при помощи метода KMeans [4]. Илюстрация кластеров точек временного ряда продемонстриирована на рис. 5.

5 Заключение

В работе рассматривалась задача поиска характерных переодических структур внутри временого ряда. Рассматривался метод основаный на локальном снижение размерности фазового пространства. Был предложен алгоритм поиска, который основывается на методе главных компонент для локального снижения размерности, а также на использовании некоторой функции растояния между локальными базисами в каждый момент времени, которые интерпретировались как признаковое описание точки временного ряда.

В ходе эксперимента было показано, что предложеный метод измерение расстояния между базисами хорошо разделяет точки которые принадлежат различным классам, что приводит к хорошей кластеризации объектов.

Предложеный метод имеет ряд недостаткров связаных с большим количеством ограничей на временной ряд. Данные ограничения будут ослаблены в последнующих работах.

Список литературы

[1] *И. П. Ивкин, М. П. Кузнецов* Алгоритм классиификации временных рядов акселерометра по комбинированному признаковому описанию. // Машинное обучение и анализ даннызб 2015.

[2] *V. V. Strijov, A. M. Katrutsa* Stresstes procedures for fetures selection algorithms. // Schemometrics and Intelligent Laboratory System, 2015.

[3] *I. Borg, P. J. F. Groenen* Modern Multidimensional Scaling. — New York: Springer, 2005. 540 p.

[4] *T. Kanungo, D. M. Mount et al* An Efficient k-Means Clustering Algorithm: Analysis and Implementation. 2000.