Восстановление прогноза, сделанного в метрическом вероятностном пространстве, в исходное пространство (временных рядов)

A Preprint

Maxim Divilkovskiy
Chair of Data Analysis
MIPT
divilkovskii.mm@phystech.edu

Vadim Strijov
FRC CSC of the RAS
Moscow, Russia
strijov@phystech.edu

Abstract

Исследование посвящено задаче прогнозирования набора временных рядов с высокой ковариацией. Исследуются наборы временных рядов с высокой дисперсией. Для решения данной задачи предлагается построение пространства парных расстояний, представляющего метрическую конфигурацию временных рядов. Прогноз осуществляется в этом пространстве, а затем результат возвращается в исходное пространство. В данной статье рассматриваются методы перевода прогноза из метрического пространства в исходное пространство временных рядов. Помимо этого, приводится оценка качества прогноза. Новизна работы заключается в использовании только метрического пространства для прогноза.

Keywords Riemannian Space · Trades · Multidimensional Scaling · Time Series

1 Introduction

Временные ряды возникают во многих прикладных задачах, таких как анализ физической активности, мозговых волн или биржевых котировок. Цель данной работы заключается в представлении нового метода прогнозирования временных рядов, характеризующихся высокой попарной ковариацией. Задача разбивается на три этапа: сначала исходное пространство временных рядов некоторым образом трансформируется в метрическое пространство, затем в этом пространстве производится прогноз матрицы попарных расстояний, после чего результат возвращается в исходное пространство. В данной статье исследуется восстановление ответа в пространство временных рядов.

Классические способы предсказания временных рядов, такие как LSTM [3], SSA [2] и многие другие [6], [1] основаны на предсказании значения одного ряда, тогда как в данной работе предлагается анализировать изменение набора временных рядов. Подобное исследование проводится в статье [4], однако в ней делается упор на задаче feature selection.

Новизна работы заключается в том, что прогнозирование делается не в исходном пространстве, а в пространстве попарных расстояний. Преимущество данного метода заключается в том, что на реальных наборах временных рядов часто наблюдается зависимость, близкая к линейной, и эта дополнительная информация может улучшить качество итогового прогноза. Помимо этого, прогнозируемую матрицу можно рассматривать как набор временных рядов, однако в этом случае размерность данных возрастает до $O(n^2)$ против n рядов, что увеличивает информативность входных данных.

Далее будут рассматриваться условия, при которых можно выбрать функцию расстояния между рядами. Основными критериями выбора являются информативность расстояния, а так же возможность восстановить прогноз в пространство временных рядов.

Эксперимент проводится на синтетических, природных и финансовых временных рядах. Цель эксперимента заключается в выборе наилучшего способа построения метрического пространства.

2 Problem Statement

2.1 Formal Problem

Предполагается, что набор временных из d рядов задан t векторами:

$$[\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_t], \forall k : \vec{x}_k \in \mathbb{R}^d$$

$\vec{x}_{t_i,k}$ задаёт собой значение ряда с индексом k в момент времени t_i.

Задача заключается в прогнозе $\vec{x}_{t+1}$.

В дальнейшем, $\vec{x}$ будем так же называть *многомерным* временным рядом, рассматривая значение в точке как элемент пространства $\mathbb{R}^d$.

2.2 Algorithm

1. Строятся матрицы расстояний по предыдущим шагам.

$$\begin{gathered}[\vec{x}_1, \ldots, \vec{x}_s] \to \Sigma_s \\ [\vec{x}_2, \ldots, \vec{x}_{s+1}] \to \Sigma_{s+1} \\ \vdots \\ [\vec{x}_{t-s}, \ldots, \vec{x}_t] \to \Sigma_t\end{gathered}$$

2. По этим матрицам прогнозируется матрица $\hat{\Sigma}_{t+1}$

3. Найти такой $\hat{x}_{t+1}$, что $||\hat{\Sigma}_{t+1} - \bar{\Sigma}_{t+1}||_2^2$ минимальна, где $\bar{\Sigma}_{t+1}$ — матрица расстояний, построенная по набору $[\vec{x}_{t-s+1}, \ldots, \hat{x}_{t+1}]$.

3 Метрика

При условии высокой попарной корреляции входных рядов и постановке задачи о предсказании значения рядов в следующий момент времени необходимо определить достаточные входные данные для модели.

В данном параграфе рассматривается возвращение прогноза из матрицы Σ_{t+1} в пространство временных рядов в предположении, что матрица предсказана идеально.

Недостаточность одной матрицы попарных расстояний

Пусть дана предсказанная матрица попарных расстояний Σ размера $d \times d$ для многомерного временного ряда $\overline{X} \in \mathbb{R}^{d \times t}$. Предсказывается $y \in \mathbb{R}^d$. Так же, известна метрика $d : \mathbb{R}^{t+1} \times \mathbb{R}^{t+1} \to \mathbb{R}$, введённая на временных рядах, обладающая свойствами метрики. То есть, $\Sigma_{i,j} = d(X_i \circ y_i, X_j \circ y_j)$, где $\circ$ означает конкатенацию векторов.

В качестве примера рассмотрим евклидову метрику:

$$d(p, q) = \sqrt{\sum_{k=1}^{n} (p_k - q_k)^2}.$$

Использование данной метрики приводит к тому, что прибавление ко всем координатам y некоторой константы C не изменяет ответ. В случае задачи предсказывания временных рядов это свойство критично, поскольку даже в случае верного предсказания матрицы Σ невозможно понять как себя поведут временные ряды в момент времени $t + 1$.

Это приводит к невозможности использования алгоритма MDS для восстановления ответа в исходное пространство временных рядов.

Однако, даже использование других метрик не позволяет избавиться от проблемы.

Теорема 1. *Для любой метрики, введённой в пространстве временных рядов $\mathbb{R}^t$, существует более одного способа восстановить исходные временные ряды по построенной матрице попарных расстояний.*

Доказательство. Достаточно показать, что метрика не является биекцией. Это будет означать, что существуют несколько различных пар рядов, расстояние между которыми одинаковое.

Покажем, что метрика — непрерывная функция. Возьмём последовательность $\{(x_n, y_n)\} \subset \mathbb{R}^t \times \mathbb{R}^t, (x_n, y_n) \to (x, y)$. Тогда, $x_n \to x, y_n \to y \Rightarrow d(x_n, x) \to 0, d(y_n, y) \to 0$ при $n \to \infty$. Воспользовавшись неравенством треугольника для метрики, получаем $d(x_n, y_n) \leq d(x_n, x) + d(x, y) + d(y_n, y) \to d(x, y)$, следовательно, $d(x_n, y_n) \to d(x, y)$.

То есть метрика — непрерывное отображение из $\mathbb{R}^t \times \mathbb{R}^t$ в $\mathbb{R}$. Покажем, что такое отображение не может быть гомеоморфизмом. Предположим, что $f : \mathbb{R} \to \mathbb{R}^t \times \mathbb{R}^t$ — искомый гомеоморфизм. Возьмём некоторую точку $a \in \mathbb{R}$ и $f(a)$. Выкинув точку a, $\mathbb{R}$ перестаёт быть связным, а $\mathbb{R}^t \times \mathbb{R}^t$ нет. Значит, это не гомеоморфизм. Противоречие. ■

Замечание. Существенно, в доказательстве используется только непрерывность функции. Это означает, что даже не метрические функции не дадут единственность ответа. Например, попарная корреляция рядов тоже является непрерывной функцией.

Исходя из этих утверждений, использование одной матрицы расстояний не позволяет решить задачу прогнозирования.

4 Попарная корреляция

4.1 Построение матрицы

В данной секции рассматривается способ восстановления прогноза, использующий несколько матриц.

Матрица попарных расстояний строится следующим образом:

$$\Sigma_T = \frac{1}{T}\sum_{t=1}^{T}(x_t - \mu_T)(x_t - \mu_T)^T$$

$$\mu_T = \frac{1}{T}\sum_{t=1}^{T} x_t$$

Предположим, что мы идеально спрогнозировали матрицу таким образом. Функция $||\hat{\Sigma}_{t+1} - \bar{\Sigma}_{t+1}||_2^2$ будет иметь два минимума, задающихся явно следующим образом **TODO ДОБАВИТЬ ВЫВОД ФОРМУЛ**:

$$\hat{y_i} = y_i$$
$$\hat{y_i} = \frac{2}{T-1}\sum_{k=1}^{T-1} a_{ik} - y_i,$$

где $\hat{y_i}$ — i-я координата предсказываемого значения ряда в момент $T+1$, $A = (a_{ik})$ — исходный многомерных временной ряд, y_i — истинные значения ряда в момент $T+1$

4.2 Алгоритм прогноза

Представленные выше алгоритм возвращает два ответа, из которых невозможно выбрать нужный.

Предлагается следующий алгоритм:

1. Зафиксируем T и $T' : T \neq T'$.

2. Для T и T' произведем полученный выше алгоритм и получим наборы ответов: $[ans_1, ans_2], [ans'_1, ans'_2]$.

3. Найдём тот ответ, который лежит в пересечении.

4.3 Алгоритм при неидеальном прогнозе

Проблема вышеописанного алгоритма заключается в том, что при неидеальном прогнозе, пересечения может не быть. Более того, если брать два ответа с минимальным расстоянием, мы можем взять не ту пару ответов.

Вместо двух значений T и T' предлагается брать K значений.

Тогда мы молучим K наборов ответов:

$$[ans_{11}, ans_{12}],\ [ans_{21}, ans_{22}],\ \vdots [ans_{K1}, ans_{K2}],$$

Далее предлагается перебрать 2^K наборов ответов и выбрать тот набор, в котором диаметр минимален.

TODO Математическая логика

5 Computational Experiment

Исследуются следующие алгоритмы прогнозирования:

- LSTM [3]
- SARIMA [7]
- LSTM на матрице расстояний
- Bidirectional LSTM на матрице расстояний

5.1 LSTM

LSTM, в отличии от обыкновенной RNN позволяет выделять как кратковременные, так и долговременные зависимости, что позволяет с довольно высокой точностью прогнозировать временные ряды.

В качестве теста используется зашумленный временной ряд длины T, состоящий из суммы синусов и косинусов разных амплитуд и сдвигов. Из этого временного ряда генерируется выборка следующим алгоритмом:
1. Выбирается размер окна W.
2. Ряд разбивается на $T - W - 1$ окон размера $W + 1$ со сдвигом 1. Эти окна будут семплами
3. В каждом из полученных окон первые W будут аргументами на данном семпле, а последнее — результатом.

Ряд восстанавливается неплохо, однако минусом является то, что при усложнении данных сильно растёт сложность модели. Так же, LSTM не может работать с многомерными рядами.

5.2 SARIMA

ARIMA позволяет находить авторегрессионные зависимости. SARIMA (Seasonal ARIMA) учитывает так же сезонность данных. Это может быть полезным в случае с данными природного характера, как например, температура воздуха или выработка электричества.

Ряд прогнозируется довольно плохо, в случае если он имеет достаточно нетривиальную структуру. Так же, в данных может не быть явной сезонности, что ухудшает точность данного метода.

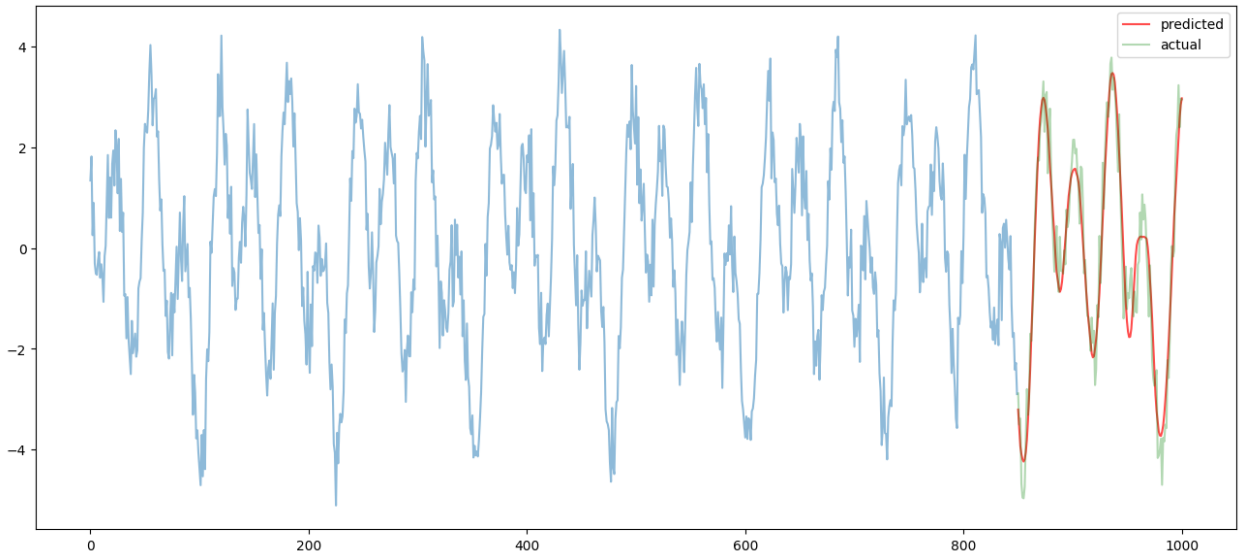

Рис. 1: Прогноз с использованием LSTM

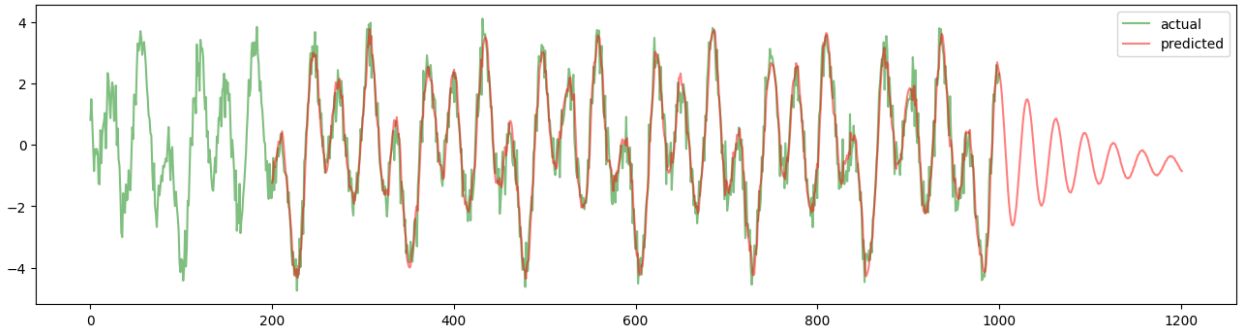

Рис. 2: Прогноз с использованием SARIMA

5.3 LSTM на матрице попарных корреляций

5.4 Bidirectional LSTM на матрице попарных корреляций

Список литературы

[1] Stephen Boyd, Enzo Busseti, Steven Diamond, Ronald N. Kahn, Kwangmoo Koh, Peter Nystrup, and Jan Speth. Multi-period trading via convex optimization, 2017.

[2] James B. Elsner and Anastasios A. Tsonis. Singular spectrum analysis: A new tool in time series analysis. 1996.

[3] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9:1735–80, 12 1997.

[4] Roman Isachenko and Vadim Strijov. Quadratic programming feature selection for multicorrelated signal decoding with partial least squares. *Expert Systems with Applications*, 207:117967, 11 2022.

[5] Shengsheng Lin, Weiwei Lin, Wentai Wu, Feiyu Zhao, Ruichao Mo, and Haotong Zhang. Segrnn: Segment recurrent neural network for long-term time series forecasting, 2023.

[6] Anastasia Motrenko and Vadim Strijov. Extracting fundamental periods to segment biomedical signals. *IEEE journal of biomedical and health informatics*, 20, 08 2015.

[7] Sima Siami-Namini and Akbar Siami Namin. Forecasting economics and financial time series: Arima vs. lstm, 2018.

[8] Ailing Zeng, Muxi Chen, Lei Zhang, and Qiang Xu. Are transformers effective for time series forecasting?, 2022.

[9] Haoyi Zhou, Shanghang Zhang, Jieqi Peng, Shuai Zhang, Jianxin Li, Hui Xiong, and Wancai Zhang. Informer: Beyond efficient transformer for long sequence time-series forecasting, 2021.