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 проблеме прогнозирования временных рядов с высокой ковариацией. Задача решается для наборов временных рядов с высокой дисперсией, проявляющейся, например, в сигналах головного мозга или ценах финансовых активов. 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 риманова пространства в качестве метрического, а также в использовании римановых моделей.

Keywords Riemannian Space · Trades · Multidimensional Scaling · Time Series

1 Introduction TO BE REWRITTEN

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 для конкретного типа временных рядов, характеризующихся высокой дисперсией и высокой попарной ковариацией. Задача разбивается на три этапа: сначала исходное пространство временных рядов трансформируется в метрическое пространство (по попарным расстояниям), затем в этом пространстве производится прогноз, после чего результат возвращается в исходное пространство. В данной статье исследуется восстановление ответа в пространство временных рядов, то есть третий этап задачи. Также проводится оценка качества прогноза.

Классические способы предсказания временных рядов, такие как LSTM [3], SSA [2] и многие другие [5], [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.

Метрическое пространство выбирается таким образом, чтобы из него можно было получить ответ. Помимо попарных скалярных произведений, можно использовать функции, являющиеся *ядрами*, то есть удовлетворяющие условиям Мёрсера.

Эксперимент проводится на биологических и финансовых данных. Цель эксперимента заключается в выборе наилучшего способа построения метрического пространства.

2 Problem Statement TO BE REWRITTEN

2.1 Formal Problem

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

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

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

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

2.2 Base Algorithm

1. Построить матрицу расстояний.

$$\hat{\Sigma}_T = \frac{1}{T}\sum_{t=1}^{T}(x_t - \mu_T)(x_t - \mu_T)^T \tag{1}$$

$$\mu_T = \frac{1}{T}\sum_{t=1}^{T} x_t \tag{2}$$

2. Спрогнозировать матрицу расстояний на следующем моменте времени $\hat{\Sigma}^s_{T+1} \approx \hat{\Sigma}_{T+1}|\hat{\Sigma}_T$. Линейная регрессия:

$$\hat{\Sigma}^s_{T+1} = W \cdot \hat{\Sigma}_T \tag{3}$$

3. Найти такой оптимальный x_{T+1}, что ошибка прогнозирования временных рядов минимальна.

3 Computational Experiment

Исследуются следующие алгоритмы прогнозирования:

- LSTM [3]
- SARIMA [6]
- MSSA [2]

3.1 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.

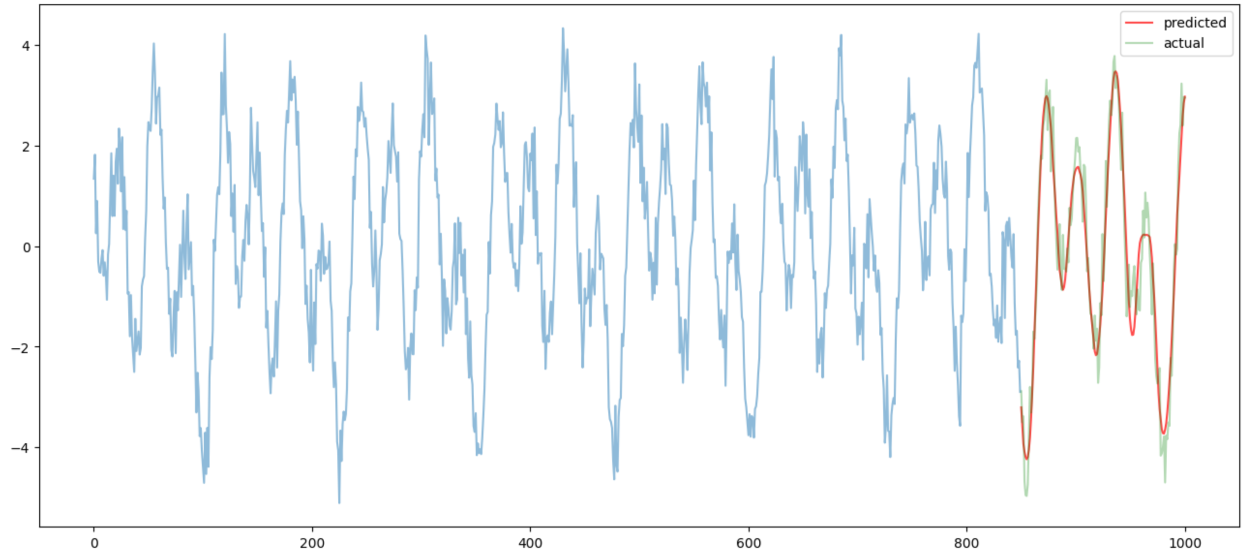

Рис. 1: Прогноз с использованием LSTM

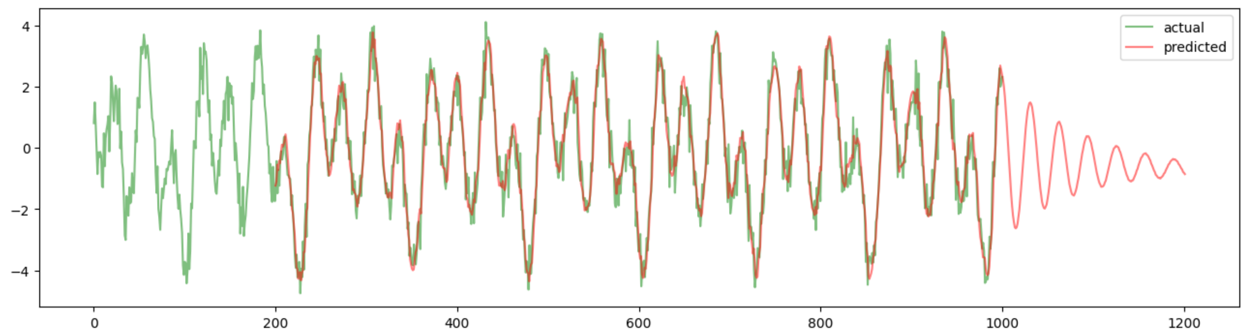

Рис. 2: Прогноз с использованием SARIMA

3.2 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.

3.3 MSSA

MSSA (Multivariate Singular Spectrum Analysis) в отличие от других методов позволяет брать во внимание корреляцию между несколькими рядами и прогнозировать несколько.

4 Метрика

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

Недостаточность 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. Рассмотрим метрику $d(x,y)$ как функцию из $\mathbb{R}^{2t+2}$ в $\mathbb{R}$. Известно, что метрика не сюрьективна, поскольку $d(x,x) = 0$. Из этого следует существование нескольких возможных ответов на задачу. **TODO: возможно стоит использовать теорему Рисса об эквивалентности норм в конечномерных пространствах.**

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

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

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