Погружение временных рядов с высокой волатильностью в метрическое пространство

Алтай Эйнуллаев Эльшан оглы

Московский физико-технический институт

Курс: Автоматизация научных исследований (практика, В. В. Стрижов)/Группа 105

Эксперт: д.ф-м.н. В.В.Стрижов

Консультант: К.Яковлев

2024

Цель исследования

Решается задача прогнозирования набора сильно коррелированных временных рядов с высокой волатильностью. Предложен метод COF (correlation optimization forecasting) использующий матрицу попарных расстояний между временными рядами набора. Исследуются влияние различных способов вычисления попарного расстояния на точность прогноза и сравнивается точность прогноза предложенного метода с ARIMA и LSTM.

Постановка задачи

Пусть $\mathrm{X} = \{\mathbf{x} = [x_1, \ldots, x_\mathrm{m}]^T | x_i \in R\}$ — множество временных рядов, заданных своей реализацией. Обозначим через $\mathbf{Y} \in R^{n \times m}$ заданный набор из n временных рядов:

$$\mathbf{Y} = [\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)}]^T. \tag{1}$$

Через $\mathbf{Y}_t \in R^{n \times t}$ обозначим $t < m$ первых столбцов $\mathbf{Y}$:

$$\mathbf{Y}_t = [\mathbf{x}_{1:t}^{(1)}, \ldots, \mathbf{x}_{1:t}^{(n)}]^T. \tag{2}$$

.

Таким образом, по известной $\mathbf{Y}_t$ выполняется прогноз значений набора временных рядов в момент времени $t + 1$:

$$\mathbf{Y}_t \to \mathbf{x}_{t+1} = [x_{t+1}^{(1)}, \ldots, x_{t+1}^{(n)}]^T. \tag{3}$$

.

Решение

Определим функцию расстояния между временными рядами: $d : \mathrm{X} \times \mathrm{X} \to R$, удовлетворяющую условиям Мерсера.

$$d_t(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}) = d_t(i, j) \tag{4}$$

Таким образом, в каждый момент времени t набору временных рядов $\mathbf{Y}_t$ поставлена в соответствие матрица попарных расстояний $\Sigma_t \in \mathcal{S}_n^+$ (симметричная, неотрицательно определенная матрица). С помощью ARIMA и LSTM прогнозируем матрицу попарных расстояний $\hat{\Sigma}_{t+1}$. Далее, в предположении, что временные ряды неотрицательны, возвращаем прогноз в исходное пространство:

$$\mathbf{x}_{t+1} = \arg\min_{\mathbf{x} \in R_+^n} \|\Sigma_{t+1} - \hat{\Sigma}_{t+1}\|. \tag{5}$$

Выбор функции попарных расстояний

Набор временных рядов в момент времени t:

$$\mathbf{X}_t = \begin{pmatrix} x_{t-\mathrm{L}+1}^{(1)} & x_{t-\mathrm{L}+2}^{(1)} & \dots & x_t^{(1)} \\ x_{t-\mathrm{L}+1}^{(2)} & x_{t-\mathrm{L}+2}^{(2)} & \dots & x_t^{(2)} \\ \dots & \dots & \dots & \dots \\ x_{t-\mathrm{L}+1}^{(n)} & x_{t-\mathrm{L}+2}^{(n)} & \dots & x_t^{(n)} \end{pmatrix} \tag{6}$$

i-й столбец матрицы $\mathbf{X}_t$ — $\mathbf{y}^i$, i-я строка — $\mathbf{x}^{(i)}$.

$$\Sigma_t^1 = \frac{1}{L} \sum_{i=1}^{L} \mathbf{y}^i \mathbf{y}^{iT} \tag{7}$$

Введем вектор $\mathbf{m} \in R^{\mathrm{n}}$ следующим образом:

$$\mathbf{m}_j = \frac{1}{L} \sum_{i=1}^{L} \mathbf{y}_j^i. \tag{8}$$

Выбор функции попарных расстояний

Второй способ:

$$\Sigma_t^2 = \frac{1}{L}\sum_{i=1}^{L}(\mathbf{y}^i - \mathbf{m})(\mathbf{y}^i - \mathbf{m})^T \qquad (9)$$

Для каждой строки $\mathbf{x}^{(j)}$, $j \in 1, \ldots, n$ матрицы $\mathbf{X}$ введем величину:

$$\sigma_j = \sqrt{\frac{1}{L}\sum_{i=1}^{L}(\mathbf{x}_i^{(j)} - \mathbf{m}_i)^2}. \qquad (10)$$

Теперь введем матрицу Σ_t^3 следующим образом:

$$(\Sigma_t^3)_{ij} = \frac{(\Sigma_t^2)_{ij}}{\sigma_i \sigma_j}. \qquad (11)$$

Вычислительный эксперимент

Вычислительный эксперимент проводился на наборе цен на электричество. Из набора предварительно были исключены временные ряды с выбросами. Важно, что полученный набор временных рядов удовлетворяет следующим условиям

1. Временные ряды высококоррелированны.
2. Временные ряды неотрицательны.

Первое требование следует из постановки задачи. Второе необходимо для разрешения проблемы единственности решения задачи оптимизации.

Ход эксперимента

Эксперимент состоял из четырех частей:

1. Выполнение LSTM прогноза.
2. Выполнение ARIMA прогноза.
3. Выполнение COF-LSTM прогноза для трех способов подчсета Σ.
4. Выполнение COF-ARIMA прогноза для трех способов подчсета Σ.

Для каждого метода выполнялось:

- Определение оптимального горизонта прогноза с помощью метода сломанной трости.
- Выполнение прогноза с оптимальным горизонтом прогнозирования.
- Вычисление ошибки прогноза.

Анализ ошибки

Алгоритм	MAPE	MSE
LSTM	0.15	**59.14**
COF-LSTM-1	0.46	359.67
COF-LSTM-2	0.29	222.46
COF-LSTM-3	0.43	437.85
ARIMA	**0.14**	76.62
COF-ARIMA-1	0.36	293.57
COF-ARIMA-2	0.30	225.41
COF-ARIMA-3	0.50	918.49

Таблица: Ошибки прогноза набора временных рядов различных методов.

Заключение

Результаты

- Сравнили три различных способа вычисления матрицы попарных расстояний.
- COF, при использовании LSTM, ARIMA для прогнозирования матрицы уступает простым ARIMA, LSTM.

Дальнейшие исследования

- Провести более тщательный анализ решаемой задачи оптимизации.
- Попробовать улучшить точность прогноза матрицы Σ при помощи генеративных моделей.

Литература

- Robert H Shumway, David S Stoffer, and David S Stoffer. *Time series analysis and its applications*, volume 3. Springer, 2000.
- J. Schmidhuber S. Hochreiter. , *Long short-term memory. Neural Computation* 1997.
- Ana Corber an-Vallet, Jos e D Berm udez, and Enriqueta Vercher. *Forecasting correlated time series with exponential smoothing models.*, 27(2):252–265, 2011.
- Razvan-Gabriel Cirstea, Darius-Valer Micu, Gabriel-Marcel Muresan, Chenjuan Guo, and Bin Yang. *Correlated time series forecasting using multi-task deep neural networks.*, pages 1527–1530, 2018.
- Benyamin Ghojogh, Ali Ghodsi, Fakhri Karray, and Mark Crowley (2021). *Reproducing kernel hilbert space, mercer's theorem, eigenfunctions, nyström method, and use of kernels in machine learning.*