Pattern matching for multivariate time series forecasting

ENBIS-24 Conference

Noé Lebreton/Julien Jacques/Julien Ah-Pine

16 september 2024

DATA-PRISME / ERIC Laboratory Lyon

Contents

Introduction

Explanation of the aim

For each time series i of a set I of time series, we have a time series at a set of times $\mathbf{t}_i = \{t_{i1}, \ldots, t_n\}$:

$$x_i(\mathbf{t}_i) = x_i(t_{i1}), \ldots, x_i(t_n)$$

So the objective is to produce predictions over a range of times $\mathbf{t}_h = \{t_{n+1}, \ldots, t_{n+h}\}$.

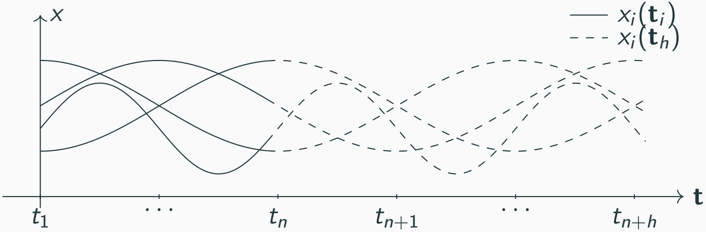


Figure 1: Representation of our problem

Weigthed Nearest Neighbors (WNN) for multivariate time series(WNN_{multi})

Weighted Nearest Neighbors (WNN) [3]

With nearest neighbor search, the objective is to predict the sequence of a curve up to the horizon h. Two hyperparameters need to be set: w and k.

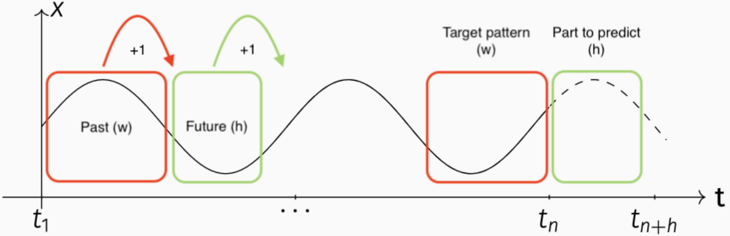
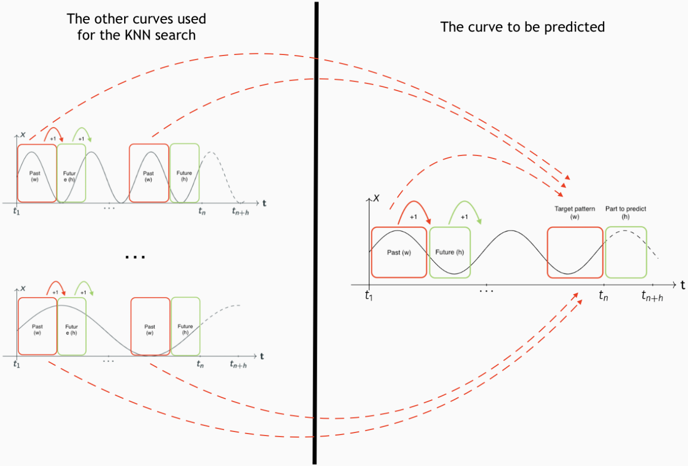


Figure 2: Past/future split using a sliding window.

Weigthed Nearest Neighbors (WNN) for multivariate time series(WNN_{multi})

In our context, we propose to use information from several curves to predict a particular curve.



Weigthed Nearest Neighbors (WNN) for multivariate time series(WNN_{multi})

WNN_{multi} **algorithm**:

- Splitting the curves to obtain 2 sets: E_p (past) and E_f (future)
- For each curve :
    - Euclidean distance between the target pattern and the other patterns in E_p.
    - Selection of k patterns.
    - Calculation of α_i weights (based on distances).
    - Calculation of the weighted average of all futures (E_f) corresponding to the k nearest neighbors.

Gaussian processes for forecasting multivariate time series

Gaussian processes

A Gaussian process is fully specified by a mean function $m(.)$ and a covariance function $K(.,.)$. For regression, we can write :

$$y_i = f_i + \epsilon_i \tag{1}$$

where $f_i \sim GP(m(.), K(.,.))$ with $m(.)$ the mean function, $K(.,.)$ the covariate function et $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$.

Multi-task Gaussian processes with common mean

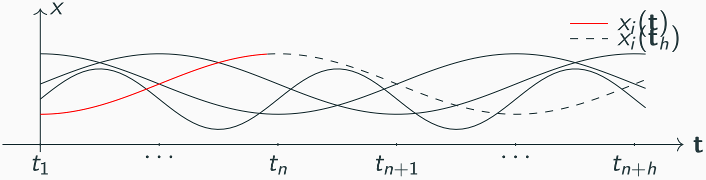


Figure 3: Representation of the multi-task Gaussian process problem

Two different approaches are proposed :

- MAGMA [1] : $y_i = \mu_0 + f_i + \epsilon_i$
- MagmaClust [2] : $y_i = \mu_k + f_i + \epsilon_i$

with $\mu_0 \sim GP(m_0(.), K_{\theta_0}(.,.))$ the mean GP common to all individuals, $\mu_k \sim GP(m_k(.), K_{\theta_k}(.,.))$ the common mean GP of the cluster k, $f_i \sim GP(m(.), K(.,.))$ the specific GP of the i-th individual et ϵ_i the noise.

MagmaKNN

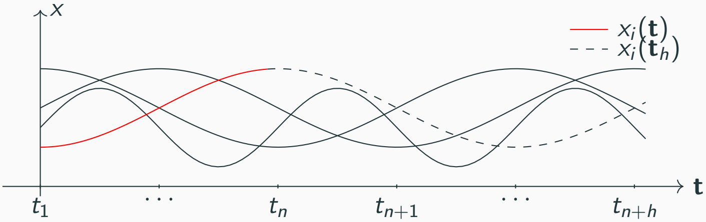


Figure 4: Representation of the multi-task Gaussian process problem

To use multi-task Gaussian processes in our context, we propose a new method MagmaKNN :

- Cutting the curves to obtain a sample of past/future sub-curves (curves in black).
- Selection of the k nearest neighbors of the target curve (curve in red).
- Using MAGMA with the reduced sample size.

Experiments

Experiments design

Data are from the electricity Spanish market (`www.omie.es`) where the objective is to predict electricity prices (expressed in Euro/MWh) and electricity demand (expressed in MW) for the next 24 hours.

- Validation dataset :
  $n = 183$ days, $k = [2, 5, 10, 20]$ and $w = [24, 48, 72, 96]$
- Test dataset : $n = 183$ days with the best hyperparameters.

Validation dataset

Data	Model	w_{best}	k_{best}	RMSE	MAE
Prices	MV-kWNN [4]	24	20	30.14/**25.07**	26.35/21.26
	WNN_{multi}	24	10	**27.63**/25.33	**23.75**/**20.34**
	MagmaKNN	24	10	28.29/24.49	24.06/20.47
Demand	MV-kWNN	24	20	1822.15/1653.6	1605.16/1407.03
	WNN_{multi}	24	20	**1717.87**/**1594.29**	**1504.94**/1359.92
	MagmaKNN	24	20	1835.06/1642.45	1605.39/**1363.5**

Table 1: Means/medians for pattern matching methods

HP for XGBoost : $n_{estimators} = [50, 100, 150]$, max depth $= [5, 10, 15]$ and learning rate $= [0.01, 0.1, 0.3]$.

HP for RF : $n_{estimators} = [50, 100, 150]$ and max depth $= [5, 10, 15]$.

Data	Model	$n_{estimators}$	max depth	learning rate (lr)	RMSE	MAE
Prices	XGBoost	100	10	0.1	**20.97**/**17.1**	**17.28**/**13.67**
	RF	150	15	-	21.78/19.01	18.1/15.27
Demand	XGBoost	50	5	0.1	**1244.01**/**1157.8**	**1024.53**/**956.04**
	RF	150	10	-	1345.86/1166.27	1092.84/966.81

Table 2: Means/medians for machine learning methods

Test dataset

Data	Model	w_{best}	k_{best}	RMSE	MAE
Prices	MV-kWNN	24	20	17.46/**13.62**	14.52/**11.78**
	WNN_{multi}	24	10	**17.2**/14.46	**14.26**/12.25
	MagmaKNN	24	10	19.89/16.17	16.64/13.51
Demand	MV-kWNN	24	20	1359.96/1198.46	1135.08/997.49
	WNN_{multi}	24	20	**1301.85**/**1107.5**	**1081.08**/**909.52**
	MagmaKNN	24	20	1416.59/1147.73	1174/925.82

Table 3: Means/medians for pattern matching methods

Data	Model	$n_{estimators}$	max depth	lr	pdq	RMSE	MAE
Prices	XGBoost	100	10	0.1	-	**15.9**/**12.31**	**12.66**/**10.06**
	RF			-	-	16.33/13.34	13.37/11.24
	ARIMAX	-	-	-	(2,1,2)	26.39/23.23	21.81/19.3
Demand	XGBoost	50	5	0.1	-	**1124.97**/**1001.82**	**931.98**/**808.26**
	RF			-	-	1229.02/1108.13	1021.22/896.32
	ARIMAX	-	-	-	(2,0,1)	5201.45/5222.88	4359.06/4264.51

Table 4: Means/medians for machine learning methods

Conclusion

Conclusion

- With our methods, we provide flexibility to the nearest neighbor search.
- The general idea of nearest neighbor selection is interesting in a multivariate context with many individuals.
- The use of machine learning methods such as XGBoost gives good results and that may be interesting for our general idea of nearest neighbor selection.

Bibliographie

A. Leroy, P. Latouche, B. Guedj, and S. Gey.
Magma: inference and prediction using multi-task gaussian processes with common mean.
Machine Learning, 111(5):1821–1849, May 2022.

A. Leroy, P. Latouche, B. Guedj, and S. Gey.
Cluster-specific predictions with multi-task gaussian processes.
Journal of Machine Learning Research, 24(5):1–49, 2023.

R. Talavera-Llames, R. Pérez-Chacón, M. Martínez Ballesteros, A. Troncoso, and F. Martínez-Álvarez.
A nearest neighbours-based algorithm for big time series data forecasting.
volume 9648, pages 174–185, 04 2016.

R. Talavera-Llames, R. Pérez-Chacón, A. Troncoso, and F. Martínez-Álvarez.
Mv-kwnn: A novel multivariate and multi-output weighted nearest neighbours algorithm for big data time series forecasting.
Neurocomputing, 353:56–73, 2019.
Recent Advancements in Hybrid Artificial Intelligence Systems.