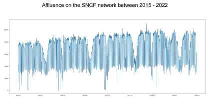
T-GCN model for forecasting of affluence in SNCF-Transilien stations

Minh-Duy NGUYEN, sous l'encadrement de Rémi COULAUD

Project context

Objective of our project is to construct a prediction model for a time serie of the affluence in each SNCF Transilien station.

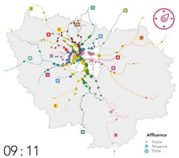


Motivation of method

One particular point of our data : **temporal** and **spatial** dependencies simultaneously.

Limitations of other classic methods :

- Not consider spatial dependencies (i.e. HA, ARIMA, SARIMA …)
- Not match the context of network topology (i.e. model only suitable for euclidean data)



09 : 11

Finding a method that takes into account both temporal and spatial dependencies is necessary for building a precise prediction model
→ **T-GCN** : temporal-graph convolutional network

Framework of method

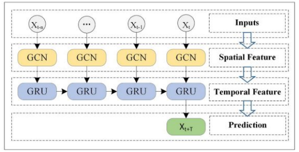


Fig. 3. Overview. We take the historical traffic information as input and obtain the finally prediction result through the Graph Convolution Network and the Gated Recurrent Units model.

The T-GCN model is composed of 2 parts :
- GCN: graph convolutional network
- GRU: gated recurrent unit

Methodology

Step 1 : Spatial dependency modelling

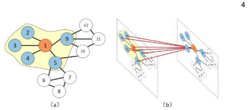


Fig. 4. Assuming that node 1 is a central road. (a) The blue nodes indicate the roads connected to the central road. (b) We obtain the spatial feature by obtaining the topological relationship between the road 1 and the surrounding roads.

model [47] to learn spatial features from traffic data. A 2-layer GCN model can be expressed as:

$$f(X, A) = \sigma\left(\widehat{A} Relu\left(\widehat{A} X W_0\right) W_1\right) \qquad (2)$$

where X represents the feature matrix, A represents the adjacency matrix, $\widehat{A} = \widetilde{D}^{-\frac{1}{2}} \widetilde{A} \widetilde{D}^{-\frac{1}{2}}$ denotes preprocessing step, $\widetilde{A} = A + I_N$ is a matrix with self-connection structure, $\widetilde{D}$ is a degree matrix, $\widetilde{D} = \sum_j \widetilde{A}_{ij}$. W_0 and W_1 represent the weight matrix in the first and second layer, and $\sigma(\cdot)$, $Relu()$ represent the activation function.

Step 2 : Temporal dependency modelling

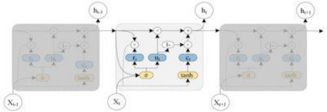


Fig. 5. The architecture of the Gated Recurrent Unit model.

$$z_t = \sigma_g(W_z x_t + U_z h_{t-1} + b_z)$$ Update gate vector

$$r_t = \sigma_g(W_r x_t + U_r h_{t-1} + b_r)$$ Reset gate vector

$$h_t = (1 - z_t) \circ h_{t-1} + z_t \circ \sigma_h(W_h x_t + U_h(r_t \circ h_{t-1}) + b_h)$$

Empirical study

Sample dataset description : **SZ Taxi**
Number of nodes : N = 156 major roads
Number of features : P = 31 days (1/1/2015 - 31/1/2015)

For the phase of model validation, we use these **metrics**

(1) Root Mean Squared Error (RMSE):

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_t - \widehat{Y}_t)^2}$$

(2) Mean Absolute Error (MAE):

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|Y_t - \widehat{Y}_t\right|$$

(3) Accuracy:

$$Accuracy = 1 - \frac{\| Y - \widehat{Y} \|_F}{\| Y \|_F}$$

(4) Coefficient of Determination (R2):

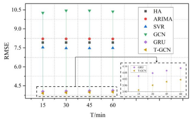
$$R^2 = 1 - \frac{\sum_{i=1}(Y_t - \widehat{Y}_t)^2}{\sum_{i=1}(Y_t - \bar{Y})^2}$$

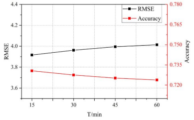
(5) Explained Variance Score (Var):

$$var = 1 - \frac{Var\left\{Y - \widehat{Y}\right\}}{Var\left\{Y\right\}}$$

Results on sample dataset

Prediction performance





T	Metric	SZ-taxi					
		HA	ARIMA	SVR	GCN	GRU	T-GCN
15min	$RMSE$	7.9198	8.2151	7.5368	9.2717	4.0483	3.9162
	MAE	5.4969	6.2192	4.9269	7.2606	2.6814	2.7061
	$Accuracy$	0.6807	0.4278	0.6961	0.6433	0.7178	0.7306
	R^2	0.7914	0.0842	0.8111	0.6147	0.8498	0.8541
	var	0.7914	*	0.8121	0.6147	0.8499	0.8626
30min	$RMSE$	7.9198	8.2123	7.4747	9.3450	4.0769	3.9617
	MAE	5.4969	6.2144	4.9819	7.3211	2.7009	2.7452
	$Accuracy$	0.6807	0.4281	0.6987	0.6405	0.7158	0.7275
	R^2	0.7914	0.0834	0.8142	0.6086	0.8477	0.8523
	var	0.7914	*	0.8144	0.6086	0.8477	0.8523
45min	$RMSE$	7.9198	8.2132	7.4755	9.4023	4.1002	3.9950
	MAE	5.4969	6.2154	5.0332	7.3704	2.7207	2.7666
	$Accuracy$	0.6807	0.4280	0.6986	0.6383	0.7142	0.7252
	R^2	0.7914	0.0837	0.8141	0.6038	0.8460	0.8509
	var	0.7914	*	0.8142	0.6039	0.8459	0.8509
60min	$RMSE$	7.9198	8.2063	7.4883	9.4504	4.1241	4.0141
	MAE	5.4969	6.2118	5.0714	7.4120	2.7431	2.7889
	$Accuracy$	0.6807	0.4282	0.6981	0.6365	0.7125	0.7238
	R^2	0.7914	0.0825	0.8135	0.5998	0.8442	0.8503
	var	0.7914	*	0.8136	0.5999	0.8321	0.8504

Proposition for application on real SNCF dataset

For applying this model on our real dataset, it is necessary to
- Re-preprocess the SNCF dataset in order to adapt to the two-part form, i.e, to create an adjacency matrix A that contains only 0 and 1. If 2 stations are linked to each other by a train-line, we put 1 in respective case, 0 otherwise.
- Clean and transform the data to adapt to the model input.

Conclusion

We presented a novel neural network-based approach for railway affluence forecasting, named T-GCN, which can solve both problems of temporal and spatial dependencies in traffic time series predicting. For further work, we propose to apply other model, such as LSGCN, to tackle this dilemma.