Hybrid Machine Learning Model for Traffic Forecasting

Khezaz Abderraouf[1,2(✉)], Manolo Dulva Hina[1], Hongyu Guan[2], and Amar Ramdane-Cherif[2]

[1] ECE Paris School of Engineering, 37 quai de Grenelle, 75015 Paris, France
{abderraouf.khezaz,manolo-dulva.hina}@ece.fr
[2] Université Paris-Saclay, UVSQ, Laboratoire d'Ingénierie des Systèmes de Versailles, 78140 Vélizy-Villacoublay, France
{hongyu.guan,rca}@lisv.uvsq.fr

Abstract. Traffic prediction has been extensively studied in the past decades. Vehicle's speed is considered the main factor for traffic forecasting, but external parameters, such as the weather, can also have a strong impact. This is a case of a classification problem to which Machine Learning has shown to have strong solving potential, if trained properly. In this paper, we propose a two-level model related to traffic forecasting parameters: It is necessary that there is no missing data in the training set, then train a Neural Network able to accurately predict the traffic situation Three completion algorithms from different types (Machine learning, algebraic and statistical methods) are compared for the rebuilding of the training set. The set is then used to train a Convolutional Neural Network into predicting the state of the traffic the way a human would do. The model is evaluated on the two parts: How accurately it can complete the data set and how correct the predictions are. This work is part of the ongoing research on intelligent vehicles that are capable of determining the context of the driving environment.

Keywords: Traffic forecasting · Data augmentation · Convolutional Neural Network · K-Nearest Neighbour · Deep Learning

1 Introduction

As the number of road users increases, the number of traffic casualties does too. According to the 2020 World Health Organization (WHO) reports [1], 1.35 millions people die each year from traffic incidents, and cost most countries 3% of their gross domestic product. Noticing that those numbers have been slowly increasing over the years, the WHO response was the publication of a 60-pages documents detailing their studies and strategies in order to improve the situation. Amongst other things, they included a questionnaire for assessing the road safety situation in a country. The first component to be checked is "Data collection and systems", that addresses the availability, gathering system, quality and dealing of the data [18].

S. Paiva et al. (Eds.): SmartCity360° 2020, LNICST 372, pp. 188–199, 2021.
https://doi.org/10.1007/978-3-030-76063-2_14

This can be explained by the fact that the modern transportation environment has become a dynamic and complex network made of vehicles, infrastructure and pedestrians. This fast-changing environment compels drivers to have an acute perception of their surroundings and be focused on the event happening around them in real-time. Thanks to the surge of new technologies in Intelligent Transportation Systems (ITS), the vehicle can provide a valuable assistance to human users, and go as far as taking initiatives [8].

Fig. 1. Illustration of a traffic congestion situation. The weather, roadworks and speed limitations all contribute to the generation of a traffic jam

One of the many aspects of transportation that can be impacted by new technologies is traffic. The complex network of transportation made of vehicles does not have a determined speed-rate speed, and is more of an unpredictable and non-linear phenomenon. There is an increasing number of road users, and there is a proportional increase of risks on their lives too.

There are many unpredictable variables that can influence the state of Traffic at a given time, such as the weather or the presence of roadworks. The important amount of data generated by those factors can be too much to process by the driver alone, hence the idea of unloading this task to a smart agent embedded in the vehicle. With the computation power and the small size of intelligent components now present on cars, there is sufficient resources to build a model capable of analyzing and predicting the state of traffic in real-time. Setting up an intelligent architecture requires a reliable set of data for training and calibration, so this aspect should also be secured [9].

We propose a model that should be able to 1) Make sure the training set is accurate enough to be used for fitting and 2) A neural network that correctly

predict the traffic situation according to a set of inputs. The model is briefly illustrated in Fig. 1.

The remainder of the paper is organized as follows: First a review of traffic forecasting works is presented in Sect. 2. In Sect. 3, the traffic forecasting data processing model is discussed in depths, including the gathering of data and their reconstructions, followed by the technical details of the model building and its validation. The paper is finally concluded by an analysis and a perspective of our future works (Fig. 2).

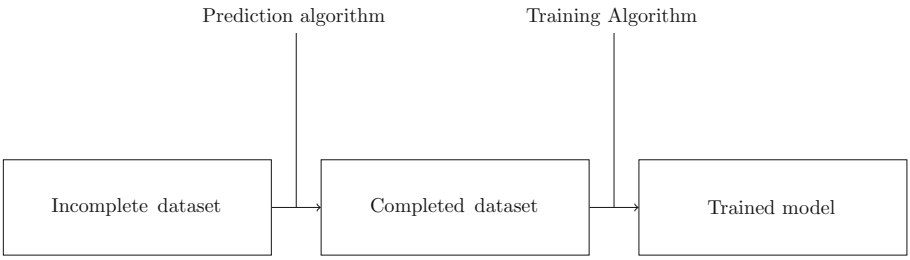

Fig. 2. A summary of the traffic forecasting model

2 Related Works

There have been many earlier works related to traffic forecasting. There have been many possible approaches to the problem; they all share the same objective – to be able to predict in the most accurate way the situation of traffic under specific conditions.

Traffic prediction has long been regarded as a statistical problem. In one of the earliest studies in 1991, Davis and Nihan [5] compared the simple univariate linear prediction a regression model in an empirical measuring of traffic congestion. They choose a Nearest-Neighbor approach and showed that this lazy-learning method was just slightly better than a classical parametric regression method. However, and as noted by the authors, the optimisation was not significant enough to be relevant, and the predictions were still not accurate enough, sometimes being up to 30% incorrect.

In 2003, with access to higher-quality traffic data, Clark [4] proposed a Non-parametric Regression model that would include other variable to speed, like the day of the week. Non-parametric regression is a form of regression that is based on the available data, rather than a pre-determined prediction function, hence being relevant in the traffic topic since there is not a single "fixed" behaviour. Their model showed great potential, but they did not have a database big enough to accurately train it, and were also lacking the computation power to properly calibrate it.

As pointed out by Vlahgioanni et al. [17], traffic forecasting has been studied for almost 3 decades now. In their literature review, they came out with 10 possible axis of improvement:

- Developing responsive algorithms and prediction schemes
- Freeway, arterial and network traffic predictions
- Short-term predictions: from volume to travel time
- Data resolution, aggregation and quality
- Using new technologies for collecting and fusing data
- Temporal characteristics and spatial dependencies
- Model selection and testing
- Compare models or combine forecasts
- Explanatory power, associations and causality
- Realizing the full potential of artificial intelligence

The last point has become the most interesting over the years. This review was made in 2014, a few months before the real surge of Deep Learning [10] and Artificial Intelligence. Many studies has since been focusing on the last improvement point proposed, i.e. building neural networks and using AI for traffic forecasting. In fact it is one of the approaches our own study is taking.

Artificial Intelligence have already been used in recent works for traffic forecasting. In 2015, Ma et al. [12] implemented a Long- short-term memory neural network to predict traffic situation. They used micro-wave detector to collect real-life data for a month and trained their model with them. They showed good results with 97% of accuracy on their predictions. Even though the model made very accurate predictions, the only input it had was the speed of cars that were going through the testing road, with no consideration of environmental data.

In 2011, Min et Wynter [14] developed a scalable method for traffic prediction up to 15 min in a dynamic environment. The mathematical model they proposed was built upon two variables: the distance and average speed of the vehicles. They showed excellent results but using only two parameters made the computation light enough to be fast. One of the possible improvements they mentioned was adding external parameters, such as "weather, incident data and roadwork, current or planned".

As stated before, most works have been focused on speed as the main traffic forecasting parameter. This paper will try to broaden the previous studies by considering a variety of other components into the prediction.

3 Traffic Forecasting Model Concept

3.1 Data Gathering

There are many types of data that can be used for traffic prediction. For this study we decided to focus on only 7 parameters and give them fixed possible values.

- **Weather**: Sunny, Cloudy or Rainy
- **Location**: City, Highway, Isolated Road
- **Day**: Weekday, Week-end
- **Time**: Rush hours, calm hours
- **Speed**: Up to 120 km/h
- **Roadwork**: Whether there are works on the road or not
- **Traffic Incident**: Whether there is a traffic incident or not

Those specific parameters were chosen after the literature review made in Sect. 2, which showed that traffic forecasting should at least include them. Data are either collected by the vehicle's sensors or the Smart City Broadcast, which is assumed to be by RF transmission of notable event. The data are also classified into 3 categories: Those related to the surrounding of the vehicles (Weather, Day and Time), the ones directly related to the car's behaviour (Speed and Location), and the events that can happen independently of the vehicle (Roadwork and traffic Incident). Those categories of data are respectively detailed in the Tables 1, 2 and 3. Based on the gathered information, we can predict the state of the traffic and categorize it into 4 different types: Light, Medium, Heavy and Extremely Heavy. The main parameters to define the output are the Roadwork and Traffic Incident variables, both being occasional and spontaneous events. They play an important role in traffic congestion, and coupled with the other parameters, such as the weather or the speed limitation, the traffic flow can be temporarily fully stopped.

Table 1. Environment parameters

Name of the parameter	Weather	Day	Time
Values of the parameter	Sunny, Cloudy or Rainy	Weekday, Week-end	Rush hours, calm hours
Gathering channel	Car sensors	Car system	Car sensors

Table 2. Car parameters

Name of the parameter	Position	Speed
Values of the parameter	City, Highway or Isolated road	up to 120 km/h
Gathering channel	Car sensors	Car sensors

Table 3. Event parameters

Name of the parameter	Roadwork	Traffic incident
Values of the parameter	Yes or No	Yes or No
Gathering channel	City broadcast	City broadcast

3.2 Data Cleaning

Once data are collected it is necessary to assure that necessary to assure that there is no missing information. For this part, three completion algorithms from different methods are compared: the algebraic SVD (Singular Value Decomposition), the statistical Mean Imputation and the learning-based classifier KNN (K-Nearest Neighbor).

- The dataset technically being an integer matrix, it is possible to use algebraic algorithms on it. One of those methods is the SVD decomposition, which has been proven to be used for matrix completion [13]
- Mean imputation is one of the easiest and most straightforward completion methods and consists of calculating the mean value of each column and using it as a replacement for the missing values. This method requires all the values to be numerical. [15]
- The KNN method takes an incomplete row of data and compares it to the most similar ones in order to predict the correct output. In this case it takes the road and traffic conditions as inputs and tries to find similar cases in the knowledge base and determine the traffic situation. [6]

These algorithms are evaluated based on the time they take to reconstruct the dataset and the accuracy of their outputs. The one with the best results is chosen as the optimal solution for the model.

3.3 Neural Network

Artificial Intelligence has become one of today's most rapidly growing technical fields, especially Machine Learning and Artificial Neural Networks (ANN). There are many different ANN architectures which efficiency depends on the task at hand, such as Recurrent Neural Networks (RNN), which requires considerable training resources but shows excellent results in time series treatment [19]. For example, one of RNN's sub-class known as Long Short-Term Memory (LSTM) [20] and using Deep-Learning methods, has been shown to even beat humans in high-level Real-Time Strategy videogames [2].

The particular case treated in this paper can be assimilated to a classification problem, to which Convolutional Neural Networks (CNN) have shown to be great solvers over the past years, in many fields such as image recognition [3] and audio recognition [7]. The second part of the model is the use of a Deep Neural Network to train the reconstructed dataset. It is expected to be able to correctly predict the output.

4 Model Implementation

4.1 Data Completion

The first step in building a model of traffic forecasting is the building of the dataset. A Python program was developed that would generate a 1000 set of

events with their respective outputs. As described in Algorithm 1, the parameters are randomly assigned an integer value that corresponds to the state of the variable. For example, Weather = 0 means that the weather is sunny and presents no problem, whereas a value of 2 means it's a rainy day that can have a strong influence on the traffic. The traffic Situation output is computed by fusing the values of Weather, Location, Day, Time, Roadwork and traffic Incident. To simulate a more realistic situation where we could have a data loss, a complementary script was added that would go through the dataset and randomly delete an information, with a chance of 5%.

Once the incomplete dataset is ready, we started the augmentation phase. Data augmentation is the process where the initial dataset is reinforced and completed. We searched for the optimal algorithm to clean it. Three methods have been considered and tested: Mean Imputation, SVD and KNN classification methods. Table 4 shows a comparison of the three algorithms performance. The KNN requires calibration in finding the value of K, and the best results seem to be reached for K = 1, meaning the algorithm replaces a missing value with the closest set resembling it. This specific case of KNN, known as the 1-nearest-neighbour, has already been shown to have excellent results for low-dimension problems [16].

Algorithm 1: Complete Set

Input : None
Output: 1000*8 Training Set

Create Training Set;
Write the headers first;
for $i \leftarrow 0$ **to** 1000 **do**
    weather = randomValueOfWeather();
    location = randomValueOfLocation();
    speed = randomValueBetween[0:120] day = randomValueOfDay();
    time = randomValueOfTime();
    trafficIncident = randomBoolean();
    roadWork = randomBoolean();
    trafficSituation = trafficComputation(weather,location,speed,day,time,
    trafficIncident,roadWork);
    set = [weather,location,speed,day,time, trafficIncident,roadWork,
    trafficSituation]

    Write the vector *set* in row i of the Training Set
end

The SVD shows overall poor performances, which is not surprising considering that it must make complex operations on the dataset. KNN and Mean Imputation both shows overall similar results with an error rate of less than 1% and the KNN having a very small advantage, but the Mean Imputation is slightly faster. The difference of speed between both algorithms is small enough to be neglected, so it is decided to go with the KNN algorithm. The dataset being local, of 7 dimensions and of relatively small size, these conditions happens to be the most advantageous for the algorithm.

Algorithm 2: Incomplete Set

Input : 1000*8 Training Set
Output: 1000*8 Incomplete training Set

```
Import the Training Set;
for i ← 0 to 1000 do
    j = randomValueBetween[0:100];
    if j > 95 then
        Random Cell from the row i = NA;
    end
end
Save the new set as Incomplete Set;
```

Table 4. Comparison between the completion algorithms.

Algorithm	Execution time (in ms)	Performance
SVD	173	32%
Mean imputation	107	99,3%
KNN	132	99,5%

4.2 Neural Network Building

The new dataset will serve for training a classification model that will then be tested with different set of scenarios. To this end, a Convolutional Neural Network was built using the Keras tool for Python. The dataset is first split into a training and testing set, according to the common 80/20 split rule. The former is used for the training of the model whilst the latter is for validation purpose. The Network is made up of 5 fully-connected layers, following the recommendations of [11] for minimizing the effect of sparsity on the fitting. For the sake of clarity and to facilitate the replication of the presented approach, the details of the CNN are listed below.

- A 512 ReLU
- A 512 Linear
- A 64 ReLU
- A 64 Linear
- A 4 Softmax layer
- To avoid over-fitting, up to 40% of dropout is introduced between the layers.

This model takes a 7-dimensions dataset (containing only the inputs) and outputs a 4-dimension vector classifying each situation to one of the possible traffic Situation values (Light, Medium, Heavy, Very Heavy).

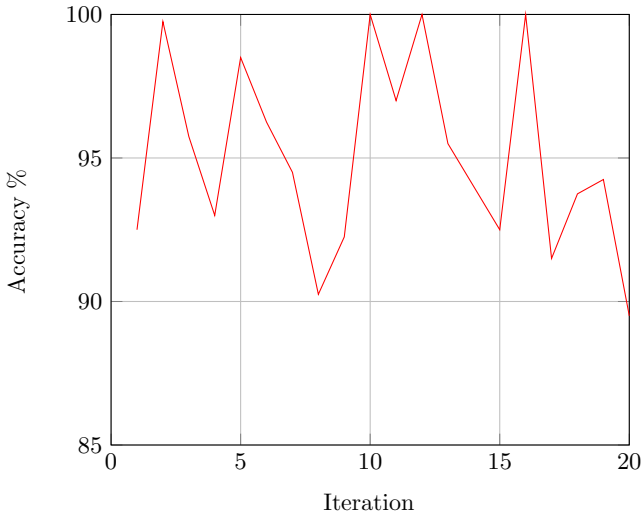

Fig. 3. Performances of the model over 20 iterations

A deep-learning architecture is described as "a multilayer stack of simple modules, all (or most) of which are subject to learning, and many of which compute non-linear input–output mappings" [10]. Our proposed network clearly matches this description, making our model a deep neural network.

In this work, the model is trained 1000 times. The fitting is done on an Ubuntu computed with an Intel *i7-8550* CPU and 16 BG of RAM memory and is made using the *ADAM* optimizer.

4.3 Model Validation

Once the CNN is built and fitted, the model is tested on different type of scenarios. Using the same algorithm as Algorithm 1, another batch of 400 sets is created and submitted to the model for classification. The intended goal is for the model to be able to predict the traffic situation the way a human would do. For example, if the weather is rainy and there is an accident on the road, a driver would automatically expect for the traffic to be heavy. On the contrary, on a sunny week-end day without any event, the traffic would be light.

When the dataset is ready, the inputs and outputs are separated, and we feed the former to the neural network. Since it is already trained, the *predict* function is used to have it classify each row of the testing set to a specific weather situation. Having saved the original outputs, they are compare with what the model computed. The tests are 20 times and the results compared every time.

As shown in Fig. 3, the model shows good prediction results, averaging around 95,03% of success with a minimum of 89% and sometimes even reaching a full 100% accuracy.

Table 5. Details of the results

Iteration	Performance in %
1	92.5
2	99.75
3	95.75
4	93
5	98.5
6	96.25
7	94.5
8	90.25
9	92.25
10	100
11	97
12	100
13	95.5
14	94
15	92.5
16	100
17	91.5
18	93.75
19	94.25
20	89.5

5 Conclusion and Future Works

In this paper, a two-level model for traffic forecasting is presented. First, an incomplete data set of road conditions is built, then compared on 3 different completion algorithms: A soft-thresholded SVD, KNN and Mean imputation. The KNN was selected for this part because of the excellent results it produced.

A deep neural network made up of 5 hidden layers was then built and trained using this dataset. Once ready, generate another set of 400 data is generated and the inputs are fed to the model. They are then compared with the previously generated outputs. This operation is repeated over 20 iterations.

The model shows overall good performances, with around 95% of correct predictions. The average time to predict output is around 10 ms (fitting time not included) (Table 5).

There are still many ways to improve our model in the future. Amongst them are the following ones:

- Training with a bigger dataset
- Challenge the KNN with more completion algorithms

– Test the model in a real-life scenario

The authors are currently building a realistic driving simulator on Unity that would allow testing the model in harsher conditions.

Another improvement point is being considered is splitting the two parts of our model in different entities: Currently everything is done in one hardware, and the idea for the future would be to have a third-party gather and complete the data (i.e. a stationary drone), then send them to the vehicle which would do the prediction according to the received information.

References

1. Road traffic injuries. https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries
2. Arulkumaran, K., Cully, A., Togelius, J.: AlphaStar: an evolutionary computation perspective. In: Proceedings of the Genetic and Evolutionary Computation Conference Companion on - GECCO 2019, pp. 314–315 (2019). https://doi.org/10.1145/3319619.3321894. http://arxiv.org/abs/1902.01724
3. Cireşan, D., Meier, U., Masci, J., Schmidhuber, J.: Multi-column deep neural network for traffic sign classification. Neural Netw. **32**, 333–338 (2012). https://doi.org/10.1016/j.neunet.2012.02.023. https://linkinghub.elsevier.com/retrieve/pii/S0893608012000524
4. Clark, S.: Traffic prediction using multivariate nonparametric regression. J. Transp. Eng. **129**(2), 161–168 (2003). https://doi.org/10.1061/(ASCE)0733-947X(2003)129:2(161)
5. Davis, G.A., Nihan, N.L.: Nonparametric regression and short-term freeway traffic forecasting. J. Transp. Eng. **117**(2), 178–188 (1991). http://ascelibrary.org/doi/10.1061/
6. Hall, P., Park, B.U., Samworth, R.J.: Choice of neighbor order in nearest-neighbor classification. Annals Stat. **36**(5), 2135–2152 (2008). https://doi.org/10.1214/07-AOS537. http://arxiv.org/abs/0810.5276
7. Hershey, S., et al.: CNN architectures for large-scale audio classification. arXiv:1609.09430 [cs, stat] (2017)
8. Ingle, S., Phute, M.: Tesla autopilot : semi autonomous driving, an uptick for future autonomy. **03**(09), 4 (2016)
9. Jordan, M.I., Mitchell, T.M.: Machine learning: trends, perspectives, and prospects. Science **349**(6245), 255–260 (2015). https://doi.org/10.1126/science.aaa8415. https://www.sciencemag.org/lookup/doi/10.1126/science.aaa8415
10. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. Nature **521**(7553), 436–444 (2015). https://doi.org/10.1038/nature14539. http://www.nature.com/articles/nature14539
11. Lin, Z., Memisevic, R., Konda, K.: How far can we go without convolution: improving fully-connected networks. arXiv:1511.02580 [cs] (2015)
12. Ma, X., Tao, Z., Wang, Y., Yu, H., Wang, Y.: Long short-term memory neural network for traffic speed prediction using remote microwave sensor data. Transp. Res. Part C Emerging Technol. **54**, 187–197 (2015). https://doi.org/10.1016/j.trc.2015.03.014. https://linkinghub.elsevier.com/retrieve/pii/S0968090X15000935
13. Mazumder, R., Hastie, T., Tibshirani, R.: Spectral regularization algorithms for learning large incomplete matrices. J. Mach. Learn. Res. **11** 2287–2322 (2010)

14. Min, W., Wynter, L.: Real-time road traffic prediction with spatio-temporal correlations. Transp. Res. Part C Emerging Technol. **19**(4), 606–616 (2011). https://doi.org/10.1016/j.trc.2010.10.002
15. Scheffer, J.: Dealing with Missing Data **3**, 8 (2002)
16. Tibshirani, S., Friedman, H.: Valerie and Patrick Hastie p. 764
17. Vlahogianni, E.I., Karlaftis, M.G., Golias, J.C.: Short-term traffic forecasting: where we are and where we're going. Transp. Res. Part C Emerging Technol. **43**, 3–19 (2014). https://doi.org/10.1016/j.trc.2014.01.005. https://linkinghub.elsevier.com/retrieve/pii/S0968090X14000096
18. World Health Organization: Save LIVES: a road safety technical package. World Health Organization, Geneva (2017)
19. Zhang, J., Man, K.F.: Time series prediction using RNN. In: Multi-dimension Embedding Phase Space p. 6
20. Zhao, Z., Chen, W., Wu, X., Chen, P.C.Y., Liu, J.: LSTM network: a deep learning approach for short-term traffic forecast. IET Intel. Transport Syst. **11**(2), 68–75 (2017). https://doi.org/10.1049/iet-its.2016.0208