DOTTORATO DI RICERCA IN INFRASTRUTTURE E TRASPORTI

SCHEDA PER L'AMMISSIONE AL II ANNO DI CORSO

Dottorando: Ken Koshy Varghese **Ciclo:** XXXVII

Curriculum: Transport And Territory Planning **Tutore:** Prof. Guido Gentile

Argomento della ricerca: Taxi ride forecast through machine learning: Anticipating demand for a better supply availability

SEZIONE A
Ricerca di Dottorato

(massimo 5 pagine)

1 – Acquisizione di conoscenze propedeutiche integrative *(contenuti appresi mediante frequenza di corsi, studio individuale, approfondimento del proprio bagaglio culturale, etc.).*

During the first year of research, the main goal was to get familiar with various machine learning techniques and models and apply them to demand forecasting problems. Therefore the following programming languages and frameworks skills were acquired:

1) Programming Languages:

- Python
- C Sharp (C#)

2) Frameworks:

- Deep Learning: Tensorflow and Keras
- Data Visualization: Matplotlib, Plotly and Seaborn
- Data Handling/Analysis: Numpy, Pandas and Scikit-learn

2 – Ricerca bibliografica svolta *(raccolta ed analisi di letteratura scientifica, con individuazione delle pubblicazioni maggiormente significative ai fini della ricerca proposta).*

The traditional approach for time series forecasting uses statistical method model for data with temporal correlation for travel demand predictions. The ARIMA model is the most extensively used method, assuming traffic forecasting is a stationary process with constant mean, variance, and auto-correlation [1]. Kalman filters [2] and Markov chains [3] are also statistical approaches. These methods usually use a linear mathematical model to determine the inner properties of the traffic flow. However, these traditional time series prediction approaches perform well in stable and linear time series prediction but struggle in non-linear and unstable time series prediction. Furthermore, they ignore the spatial correlation and work well for short-term traffic predictions.

To resolve the limitations of the traditional approach, many researchers used deep neural networks. Felix A Gers et al. [4] studied the suitability of using LSTM networks for time series forecasting. The researchers concluded that LSTM net-works could tackle the vanishing gradient problem of basic neural networks and store essential long-term information. However, regarding short term

time series forecasting, LSTM networks may not consistently outperform more straightforward strategies like ARIMA. They recommend that LSTM networks be used only after more traditional methods have failed or for long-term time series forecasting.

Using big data and machine learning, Florin Schimbinschi et al. [5] explored traffic predictions in complicated metropolitan networks. They concluded that more data leads to better ML model predictions. They also concluded that spatial relationships between road segments are a better predictor than temporal patterns. Finally, they claim that ARIMA-based models have difficulty forecasting spatio-temporal data and cannot capture complicated dynamics. Laith Alzubaidi et al. [6] is a survey paper that explains the most commonly used DL method called Convolutional Neural Network (CNN). The researcher has reviewed 300 papers published during 2010-2021 and concluded that the main benefit of CNN compared to other DL models is that it automatically identifies relevant features without human supervision. In addition, the paper describes the development of CNN architecture along with its main features, current challenges and solutions. Pedro Lara-Benítez et al. [7] investigated around seven DL methods for time series forecasting in terms of efficiency and accuracy. They trained 38000 models with a dataset having more than 50000-time series. They concluded that long short-term memory (LSTM) and convolutional neural network (CNN) are the best DL methods. LSTM obtains the most accurate forecast, and CNN performs similarly and more efficiently. It is also stated that without feature engineering requirements, CNNs can extract features from high-dimensional raw data having a grid structure, such as pixels in a picture. However, when using CNN, Huaxiu Yao et al. [8] believed that the complete city image harmed prediction accuracy. He used local CNN to introduce the semantic view and then integrated it with LSTM to improve prediction accuracy to tackle this problem and capture the spatial correlation. Even though the study captures both spatial and temporal correlation in both cases, they interacted between the two individually.

3 – Resoconto dello stato delle conoscenze relative alla tematica di ricerca *(breve sintesi del quadro scientifico di riferimento, in relazione alla tematica proposta: conoscenze consolidate e spunti per approfondimenti).*

Breif description of the literature review is provided in the section 2.

4 – Ricognizione delle attività in corso presso centri di ricerca nazionali ed internazionali *(inquadramento delle tendenze evolutive nello specifico ambito di ricerca, per quanto noto).*

Applications of AI and ML will continue to transform the corporate sector as technology develops. For instance, because of its applicability across sectors, the use of AI and ML in forecasting is of enormous importance to most businesses. In the past, businesses depended on statistical forecasting techniques like linear regressions and exponential smoothing to aid in decision-making. But across companies and sectors, machine learning-based forecasting has largely taken the role of conventional methodologies in data and analytics initiatives.

The computational requirements for ML algorithms are higher than for statistical ones. The explainability and interpretability of the models used in ML approaches is sometimes not totally understandable. However, ML approaches may be better suited for predictions in business applications with big volumes of data due to the numerous data features involved and the possibility that the algorithm utilized is not very linear or simple.

5 – Definizione della Ricerca di Dottorato *(prima formulazione del Tema per la Tesi finale, con precisazione di: finalità, metodologia, fasi e tempi delle attività previste).*

In the literature review, researches have considered the spatiotemporal correlations but minimal research has studied the effect of different spatial and time partitions. In a recent study, Liu Kai et al. [9] explored the impact of 36 spatiotemporal granularities with departure and arrival demands, revealing that a hexagonal partition with an 800 m side length and a 30 min time interval produces the best overall prediction accuracy. However, in this study, only the previous 8-time steps were considered for the next prediction. Therefore, my first objective of this research started with comparing various spatio-temporal granularity, considering a wide range of historical time steps as input features and studying its effect on the prediction performance of the deep learning models.

Proposed Methodology:

1. Data Description and Analysis to gain some insightful information that will aid in better understanding the issue.
2. Data Cleaning to remove the outliers because traffic data is acquired automatically using sensors and GPS. Therefore, databases frequently contain missing numbers, out-of-range values, incorrect temporal in-formation, and other errors. The two process done are:

a) Initial Cleaning where missing data, inaccurate pick-up or drop-off dates, unwanted columns, outbound coordinates, and sea points are removed.

b) Statistical Analysis using Z-score where outliers are defined as data points with a Z-score greater than or equal to three.

3. Spatio-Temporal 3D-Grid Clustering was done by dividing the entire city into uniform grids, and a whole day is divided into several time intervals.
4. Deep Learning Models like RNN (LSTM), CNN and Graph Neural Network (TGNet) were chosen to study the prediction performance against various spatio-temporal granularities.

The second objective of of my research is in collaboration with PTV SISTeMA, where I have been given to study and analyze two adaptive traffic signal control software called Balance and Epics.

Future steps:

1. Research on existing issues in Traffic signal management.

2. Prepare a literature review of existing research papers on machine learning in traffic signal management.

3. Study Reinforcement learning.

4. Study Data Structures and Algorithms using C#

5. Apply and analyze the efficiency of various machine learning models for adaptive traffic signal management.

6. Based on the research, develop a novel framework for adaptive control of traffic signal control with machine learning techniques.

6 – Cronoprogramma *(seguire lo schema seguente)*

n.	Attività	I Anno (consuntivo)				II Anno				III Anno			
		I	II	III	IV	I	II	III	IV	I	II	III	IV
1.	Literature review on demand forecasting	**X**	**X**	**X**									
2.	Machine learning and Deep learning fundamentals	**X**	**X**										
3.	Data collection and analysis	**X**	**X**										
4.	Spatio-Temporal 3D-Grid Clustering		**X**										
5.	Applying Deep learning models to predict Taxi demand		**X**	**X**									
6.	Explore the effect of spatio-temporal granularities		**X**	**X**									
7.	Present the first research paper in a conference			**X**									
8.	Literature review on traffic signal management				**X**	**X**	**X**						
9.	Improve the skills on data structures, algorithm and reinforcement learning					**X**	**X**						
10.	Apply and analyze the efficiency of ML models on traffic signal control						**X**	**X**					
11.	Develop a novel framework for adaptive control of traffic signal control							**X**	**X**				
12.	Evaluate the result								**X**	**X**			
13.	Write the final dissertation										**X**	**X**	**X**
14.	Present the research project in a conference												**X**
15.	Collaborate with external research groups by having an experience abroad									**X**	**X**	**X**	

SEZIONE B

Attività di collaborazione e supporto; formazione ed acquisizione di capacità evolute

(massimo 2 pagine)

1 – Partecipazione alle attività di didattica presso la struttura di afferenza *(attività seminariale, supporto alla didattica frontale, preparazione di materiale didattico, collaborazione per ricevimento studenti, collaborazione allo svolgimento di tesi di laurea e stages).*

- Assistance in Transportation and Modeling exam for Master students conducted on June 2022 and July 2022

2 – Attività di formazione *(soggiorni presso strutture di didattica e ricerca in Italia e all'estero, corsi curriculari o speciali frequentati, partecipazione a seminari, convegni, workshop, etc.).*

A. Courses: All the courses are provided by Coursera

• Machine Learning by Stanford University.

• Introduction to Machine Learning in Production

• Sequences, Time Series and Prediction

• Clustering Geolocation Data Intelligently in Python

B. Conferences:

• The 22nd International Multi-Conference Reliability and Statistics in Transportation and Communication, RelStat2020, 19-22 October 2020, Riga, Latvia

C. Seminars:

• 22 November 2021 – Prof. Guido Gentile - Programming oriented to the acquisition and cloud management of geographic data.

• 19 May 2022 – Prof. Giuseppe Loprencipe - Bibliographic database.

• 27 May 2022 – Prof. Gianluca Dell'Acqua and Prof. Salvatore Biancardo - BIM Applications.

• 28 June 2022 – Dr. Alessandro Attanasi - Introduction to Machine Learning.

• 6 July 2022 – Prof. Maria Vittoria Corazza - Preparation of international research projects.

• 23 September 2022 – Prof. Carla Nardinocchi - GIS applications.

3 – Collaborazione a studi, ricerche, programmi strutturati *(contributi in progetti di ricerca, convenzioni, etc., con inquadramento del programma e specificazione dell'attività prestata).*

SEZIONE C

Informazioni

(Tale sezione contiene le informazioni richieste alla fine ogni anno dall'Ufficio Dottorati)

1) Titolare di borsa erogata dalla Sapienza - Università di Roma…….SI☑ NO☐

2) Nazionalità …Indian……..

3) Dottorato in cotutela ……..……SI☐ NO☑

(se si indicare il cotutore……..)

4) Dottorato con doppio titolo ……..…….……SI☐ NO☑

5) Borsa con finanziamento esterno …….……SI☐ NO☑

6) Università di provenienza …"La Sapienza" University of Rome …….

7) Numero di mensilità di ricerca spese in una struttura di ricerca estera ……0……

8) Finanziamenti all'interno di reti internazionali di formazione alla ricerca ..SI☐ NO☑

9) Pubblicazioni e altri prodotti degli ultimi 3 anni

Per le aree bibliometriche. *Articoli pubblicati su riviste peer-reviewed internazionali (ed eventualmente proceedings per le aree che accettano) con impact factor (indicizzate WoS) o indicizzate Scopus.*

- Ken Koshy Varghese, Guido Gentile, et al. (2022). *Effect of Spatio-Temporal Granularity on Demand Prediction for Deep Learning Models.* 22nd International Conference on Reliability and Statistics in Transportation and Communication, RelStat'22, Riga, Latvia

Per le aree non bibliometriche. *Prodotti editoriali pubblicati dai dottorandi come Monografie dotate di ISBN e/o pubblicazioni in riviste di fascia A (o prodotti editoriali equivalenti ammessi dalla VQR).*

Bibliography

Hoel, B. W. a. L., 2003. Modeling and forecasting vehicular traffic flow as a seasonal ARIMA process: a theoretical basis and empirical results. Journal of Transportation Engi-neering, 129(6), pp. 664-672.Author, F.: Contribution title. In: 9th International Proceed-ings on Proceedings, pp. 1–2. Publisher, Location (2010).

Okutani, I., 1984. Dynamic prediction of traffic volume through Kalman filtering theory. Transportation Research Part B: Methodological, 18(1), pp. 1-11.

Qi, Y., 2014. A Hidden Markov Model for short-term prediction of traffic conditions on freeways. Transportation Research Part C: Emerging Technologies, Volume 43, pp. 95-111.

Gers, F., Eck, D. and Schmidhuber, J., 2002. Applying LSTM to Time Series Predictable Through Time-Window Approaches. Perspectives in Neural Computing, pp.193-200.

Schimbinschi, F., Nguyen, X.V., Bailey, J., Leckie, C., Vu, H., & Ramamohanarao, K. (2015). Traffic forecasting in complex urban networks: Leveraging big data and machine learning. 2015 IEEE International Conference on Big Data (Big Data), 1019-1024.

Alzubaidi, L., Zhang, J., Humaidi, A., Al-Dujaili, A., Duan, Y., Al-Shamma, O., Santama-ría, J., Fadhel, M., Al-Amidie, M. and Farhan, L., 2021. Review of deep learning: con-cepts, CNN architectures, challenges, applications, future directions. Journal of Big Data, 8(1).

Lara-Benítez, P., Carranza-García, M. and Riquelme, J., 2021. An Experimental Review on Deep Learning Architectures for Time Series Forecasting. International Journal of Neural Systems, 31(03), p.2130001.

Yao, H., Tang, X., Wei, H., Zheng, G. and Li, Z., 2019. Revisiting Spatial-Temporal Simi-larity: A Deep Learning Framework for Traffic Prediction. Proceedings of the AAAI Con-ference on Artificial Intelligence, 33, pp.5668-5675.

Liu Kai & Chen, Zhiju & Yamamoto, Toshiyuki & Tuo, Liheng. (2022). Exploring the impact of spatiotemporal granularity on the demand prediction of dynamic ride-hailing.