Traffic estimation through mobile network performance data processing

VASILEIOS ASTHENOPOULOS, IOANNIS LOUMIOTIS, PAVLOS KOSMIDES, EVGENIA ADAMOPOULOU, KONSTANTINOS DEMESTICHAS
Institute of Communication and Computer Systems
National Technical University of Athens
Iroon Polytechneiou 9, 15773, Zografou-Athens
GREECE
asthen@cn.ntua.gr

Abstract: - This paper presents the concept of using mobile network performance data in order to estimate and predict road traffic conditions. The effectiveness of the approach taken by the authors is examined using real-world data acquired from mobile and road network operators. Furthermore, a comparative analysis is performed to evaluate which of the two machine learning techniques proposed, namely the Multi-Layer Perceptron and the General Regression Neural Network is more suitable for this purpose. It is argued that practical implementations of the system described in this paper can reduce the number of sensors needed to acquire metrics from the road network, allow accurate estimation of future road traffic conditions exclusively using anonymous mobile network performance data, and even raise near real-time alerts about traffic events, without the requirement of dedicated traffic sensors.

Key-Words: - ITS, cellular data, traffic prediction, MLP, GRNN, Vehicle Detection System, mobile network

1 Introduction

Intelligent transportation systems (ITS) [1] have made significant advances in improving everyday life and reducing the cost and environmental impact of transportation. The current research trends aim for the concept of smart cities, where an entire city would coordinate and optimize every operational aspect in order to improve overall efficiency [2]. According to this approach, viewing all available data sources as potentially useful information providers is a very interesting research area.

Acknowledging that services used by citizens in everyday life (such as the cellular and road networks in our case) are not totally isolated from each other, but rather are more or less affected by one another, can lead to the conclusion that information about one system can be deduced from data provided by other systems. The study of the exact relationships among these "interlaced" systems can provide useful information about how to make use of data fusion to make reasonable assumptions about one system solely by using data originating from other systems. One simple example that helps better understand the described concept is suddenly noticing a significant increase in outgoing calls or data traffic from a specific cellular base station that is near a large highway and reaching the conclusion that there is probably some sort of traffic related incident in progress, as that would be a major reason for the drivers to stop driving and start using their mobile devices.

To this end, in [3] the authors present a survey and identify the main challenges that arise on the problem of using cellular network signalling for inferring real-time road traffic information. Similarly, a road traffic estimation system built on top of the cellular network infrastructure is presented in [4] where the authors identify the correlation between specific road conditions map and certain signalling patterns in the cellular core network.

The authors in [5] recognize the usefulness of cellular networks as alternative means for collecting road traffic information. They concentrate on collecting road traffic information from UMTS and GSM cellular systems. Finally, in [6] the authors attempt to estimate vehicular trajectories from 3G signaling traffic based on a real-world dataset from an Austrian cellular network operator. Following the above trend, this paper assesses the feasibility of using data originating from cellular networks in order to reach conclusions about traffic conditions and, more specifically, to predict vehicle speeds on an adjacent motorway.

The rest of this paper is organized as follows. In Section 2 the problem description is presented. The

used dataset is described in Section 3. In Section 4, we present the two difference approaches that we use based on machine learning. In Section 5 we provide and discuss the results that were acquired from the tests, while some future work thoughts are presented in Section 6. Finally, the paper is concluded in Section 7.

2 Problem Description

One of the most challenging tasks of road network monitoring is the management of road sensors. The cost of planning, installing and maintaining a complete set of often heterogeneous road sensors is quite significant, and can increase further, if the vast size and complexity of dense urban road networks is taken into consideration. On the other hand, in the domain of cellular communications there is the problem of creating a network dense enough to provide adequate coverage.

This paper attempts to assess the relationship between the performance figures of cellular network base stations and the traffic conditions in the adjacent road network. If close ties can be found between the two, there would be a significant benefit in utilizing data originating from the cellular network in order to analyse and/or predict road traffic and road network conditions in general. This would eventually lead to more road traffic data being indirectly available to aid in optimizing transportation, without the requirement for expensive sensor networks to be set up and maintained, but rather by simply making better use of data that is already available from the existing cellular network base stations.

3 Dataset Description

In order to assess the problem described, heterogeneous data are required, on which to validate the proposed solution. Specifically, there are two discrete data sources involved; cellular network traffic data and road traffic data. These must be properly combined in order to create a dataset which correlates the operational conditions of the cellular network with the conditions on the road network in its vicinity. The key variables of each data source used in this study is cell-to-cell handover count and road speed.

The cellular data used in this paper originate from the real-world network of a major mobile service provider in Greece, and the relevant road data are provided by Attikes Diadromes, a major Greek highway operator. In particular, a set of base stations located near key interchanges of Attiki Odos often serving increased traffic, were selected. The selection of cells was made so that their antennas point as accurately as possible towards the adjacent highway lanes. This way, it is expected that the vast majority of cellular network subscribers associated with these particular cells at any time are actually driving on the motorway.

Fig. 1 presents the alignment of the cell antennas relative to the road network, for one of the interchanges evaluated. There, it is obvious that antenna orientation is selected to follow the adjacent highway as closely as possible. Once the cells to be involved in the study were specified, the relevant road traffic data source had to be found. Attikes Diadromes operates a dense Vehicle Detection System (VDS) which is used to acquire data in real time, allowing both prompt incident response and offline statistical traffic analysis. Using their VDS, the specific induction loops that measure traffic parameters on the aforementioned interchanges were specified and monitored. Given the geospatial correlation, from the timestamps of both data sources, a temporal correlation between the two data sources can be deduced.

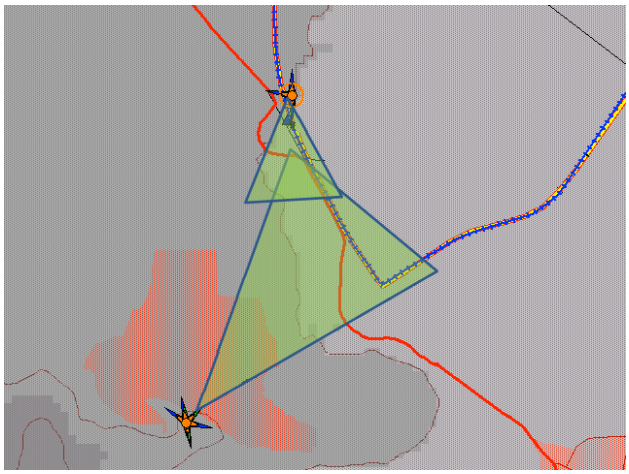

Fig. 1: Positioning of mobile network cells in relation to the road network.

The metric of interest for cellular network data is the hourly handover count. This metric refers to the number of subscribers that were disassociated from Cell A and associated (“handed over”) to Cell B within one hour. This metric can provide a quantitative figure for the directional flow of subscribers from one geographical area to the next. On the other hand, the metric taken into consideration for the road network is the average

speed of the vehicles driving on that part of the road.

In more detail, the dataset used consists of the following fields:

- *Month*; The month id. The dataset includes records for February and March 2015.
- *Day of week*; The day id (1…7).
- *Hour of day*; The hour id (0…23).
- *Cell1*; The ID of the cell the handovers originate from.
- *Cell2*; The ID of the cell the handovers are destined to.
- *Handovers*; The hourly handover count in the direction Cell1 Cell2, as measured from the cellular network.
- *Speed*; The hourly average speed over the part of road pointed to by the cells in the direction Cell1 Cell2, as measured by the VDS.

Last but not least, it is worth making clear that all data used in these experiments, from both sources, are real-world data, not simulated or expanded from a small set of real data. The full dataset used consists of about 4200 records.

4 Evaluated Techniques

The essence of the problem lies in discovering how input data (cellular network data) affect the target variable, in our case road speed. While a strict mathematical model would provide us with a well-defined relationship between these entities, a great deal of time, effort and validation would be required to successfully reach one. Therefore, a different approach is quite often used for such problems, namely machine learning techniques. The application of machine learning techniques for complex problems in recent years has proved to be quite successful, especially in areas such as text classification [7]. The Multi-Layer Perceptron (MLP) and the General Regression Neural Network (GRNN) are two very popular and proven techniques in the field of machine learning.

The Multi-Layer Perceptron [8] is one of the simplest yet most effective machine learning methods utilized these days. MLPs have proven to be successful in many different scientific fields, from simple tasks to difficult ones, such as radar target detection in noisy environments [9] or wind speed prediction [10]. Being a special case of an artificial neural network (ANN), it is designed to closely mimic the structure of the human brain and its learning abilities. An artificial neural network consists of neurons, which are formed by three elementary entities: a set of connecting links (synapses), each carrying a weight, an adder used to compute the weighted sum of the inputs carried by the synapses and an activation function which controls the output of the neuron. In an MLP, sets of neurons are structured to form layers, as depicted in Fig. 2.

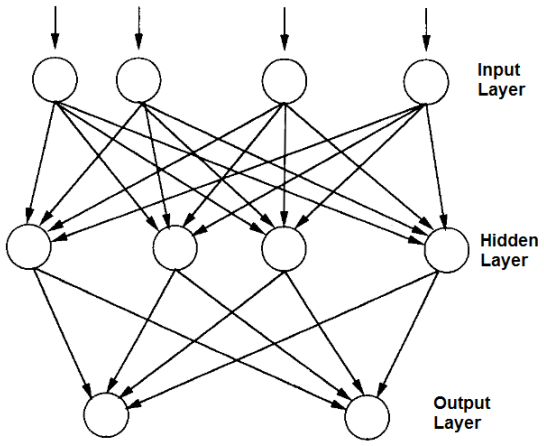

Fig. 2: A Multi-Layer Perceptron with four inputs, one hidden layer with 4 neurons and two outputs.

The first layer, called the input layer, consists of as many neurons as the number of inputs to the network, while the final, called the output layer consists of as many neurons as there are outputs. Between these layers, one or more so-called hidden layers are formed, which are fully connected with synapses to the outputs of the previous layer. Using an MLP with just one hidden layer of sufficient neurons, a large variety of functions can be approximated fairly accurately [11], while adding more hidden layers results in the system being able to mimic more complex functions. However, there is a tradeoff between the complexity of the system and the accuracy of the results it provides; increasing the number of neurons or hidden layers for a given problem could lead to worse performance and less accuracy when feeding the MLP with new, unseen data. Therefore, the architecture of the network plays a significant role in the results to be expected from it.

The General Regression Neural Network (GRNN) [12] is a single-pass neural network often utilized for the estimation of continuous variables. Its main characteristics are its quick learning ability as well as its ability to converge to the optimal regression surface as the training samples become large enough. In addition, because the regression surface is instantly defined everywhere, the GRNN

is a very suitable candidate for real-time problems that provide sparse data. One of the most challenging issues with the GRNN is the proper selection of the smoothing parameter σ that corresponds to the relevant Gaussian function used in the estimation process. In all, the structure of the GRNN is presented in Fig. 3.

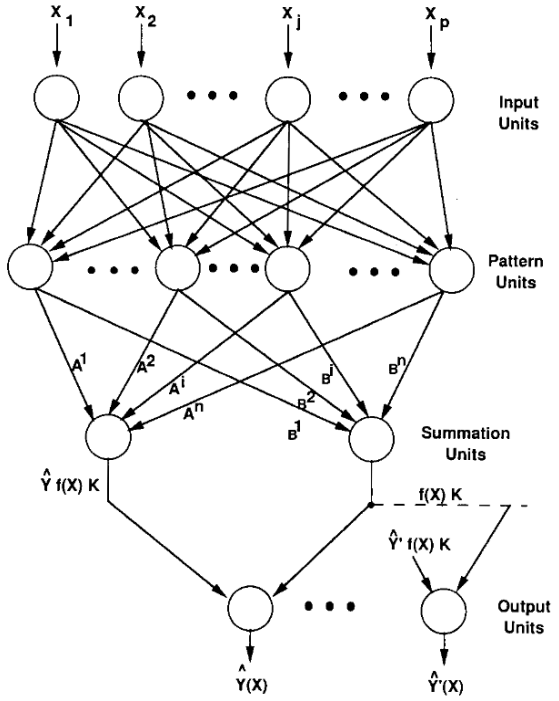

Fig. 3: Structure of a General Regression Neural Network [12].

This paper evaluates the performance of both MLPs and GRNNs for the purpose of utilizing cellular network data to predict road traffic conditions. Taking a look at the dataset, Fig. 4 depicts the relationship between hourly handover count and road speed versus the "hour of the day" variable. As can easily be seen, road speed is roughly periodic with a period of one day, while the hourly handover count presents some repetitive patters which are not as clear as those for road speed. We expect the techniques evaluated to succeed in finding the mathematical relationship between these two variables.

5 Tests and Results

The metric used to evaluate the performance of the techniques is the Mean Absolute Percentage Error (MAPE) of the target variable (road speed). For both the MLP and the GRNN, 10-fold cross validation was defined as the validation procedure. According to this popular validation procedure, the dataset is split into 10 equally sized subsets and training is performed using 9 of them as the training set and the remaining subset as the validation set. Then, the process is repeated 9 more times, rotating the subsets so that each subset is used for validation only once. This way, validation is performed on unseen data each time.

As far as the MLP run is concerned, the use of just 1 hidden layer was selected formulating a 6-7-1 three-layered feed-forward neural network, while the Logistic function was used as an activation function for the hidden layer and Linear for the output layer. Regarding the GRNN, the optimal sigma value (σ) was selected for each input variable and a Gaussian kernel function was defined. The results of the evaluations are presented in the following table.

Table I: Accuracy scored by evaluated techniques.

Technique	MAPE
GRNN	4.47%
MLP (6-7-1)	9.32%

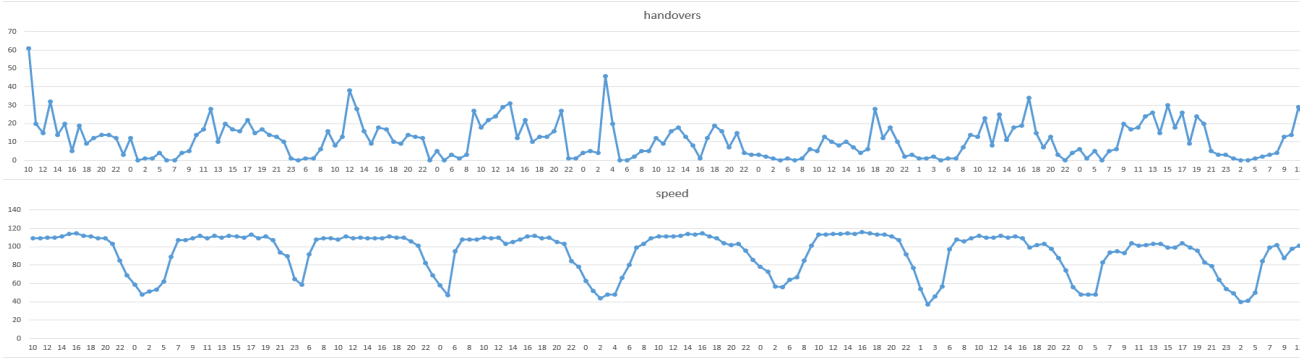

Fig. 4: Temporal relationship between hourly handover count and hourly average speed.

As can be seen from the results, the GRNN has a significantly smaller MAPE than the MLP. This shows that for this particular problem, the GRNN approach is much more appropriate. Furthermore, this level of accuracy (about 4.5%) is high enough to show that over-fitting to this specific dataset did not occur, but at the same time low enough to present usable results. This leads to the conclusion that handover data can indeed be associated with traffic conditions on the nearby road network. As a result, this kind of data can successfully be used to assess and predict road traffic parameters.

6 Future Work

This paper evaluated the relationship between two specific variables, namely the hourly handover count from the cellular network and the average hourly speed on the motorway beside the base stations considered. It would be very interesting to also take other variables, such as road traffic flow and density and data traffic or outgoing call rate into consideration for the same purposes. Furthermore, fusion of data from more than one cellular network operator could potentially provide more accurate results or better validation of the model. Also, other cellular network metrics, such as data traffic could be evaluated to assess their effect on the accuracy of predictions. Finally, an expansion of the research to more complex road networks than just the adjacent highway, potentially also evaluating more machine learning techniques, would provide a generalization of the results for urban road networks which often lack road sensors. This would make the proposed system very useful for practical implementations and lead the way to its widespread use.

7 Conclusion

This paper evaluated the concept of using cellular network performance data for the purpose of estimating and predicting road traffic parameters. Two popular machine learning techniques, namely the Multi-Layer Perceptron and the General Regression Neural Network were used to test the feasibility and accuracy of the proposed system. The results of applying these techniques to a real-world dataset showed that the GRNN provides significantly better accuracy in predicting road speed than the MLP. Further work in this research area could greatly decrease the number of road sensors needed to gain a complete view of road traffic conditions in an urban network by utilizing performance data collected from cellular network base stations.

Acknowledgment:
This work has been performed under the Greek National project CARMA (11ΣΥΝ_10_877), which has received research funding from the Operational Programme “Competitiveness & Entrepreneurship” of the National Strategic Reference Framework NSRF 2007-2013. This paper reflects only the authors’ views, and the Operational Programme is not liable for any use that may be made of the information contained therein.

References:

[1] A. Sheng-hai et al., A Survey of Intelligent Transportation Systems, Proceedings of the 3rd International Conference on Computational Intelligence, Communication Systems and Networks, Bali, July 2011, pp. 332-337.

[2] M. Brenna, M.C. Falvo, F. Foiadelli, L. Martirano, F. Massaro, D. Poli, A. Vaccaro, Challenges in energy systems for the smart-cities of the future, Proceedings of the 2012 IEEE International Energy Conference and Exhibition (ENERGYCON), 9-12 Sept. 2012, pp. 755-762.

[3] D. Valerio, A. D’Alconzo, F. Ricciato, W. Wiedermann, Exploiting Cellular Networks for Road Traffic Estimation: A Survey and a Research Roadmap, Proceedings of the IEEE 69th Vehicular Technology Conference (VTC), 26-29 April 2009, Barcelona.

[4] D. Valerio, T. Witek, F. Ricciato, R. Pilz, W. Wiedermann, Road traffic estimation from cellular network monitoring: A hands-on investigation, Proceedings of the IEEE 20th International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC), 13-16 Sept. 2009, Tokyo

[5] David Gundlegård and Johan M Karlsson, Generating Road Traffic Information from Cellular Networks - New Possibilities in UMTS, Proceedings of the 2006 6th International Conference on ITS Telecommunications, 1128-1133.

[6] P. Fiadino, D. Valerio, F. Ricciato, K.A. Hummel, Steps towards the Extraction of Vehicular Mobility Patterns from 3G Signaling Data, Proceeding of the 4th International Workshop on Traffic Monitoring and Analysis, 12 March 2012, Vienna, Austria

[7] M. Ikonomakis, S. Kotsiantis, V. Tampakas, Text Classification Using Machine Learning Techniques, WSEAS Transactions on Computers, Issue 8, Volume 4, 2005, pp. 966-974.
[8] S. Haykin, Neural Networks, A Comprehensive Foundation, Prentice Hall, 2nd edition, 1999.
[9] D. de la Mata-Moya, P. Jarabo-Amores, R. Vicen-Bueno, L. Cuadra-Rodriguez, and F. Lopez-Ferreras, MLPs for detecting radar targets in gaussian clutter, Proceedings of the 5th WSEAS International Conference on Artificial Intelligence, Knowledge Engineering and Data Bases (AIKED'06), USA, 2006, 259-264.
[10] P.M. Fonte, Goncalo Xufre Silva, J.C. Quadrado, Wind Speed Prediction using Artificial Neural Networks, Proceedings of the 6th WSEAS International Conference on Neural Networks, Portugal, 2005, pp.134-139.
[11] T.M. Mitchell, Machine Learning, McGraw-Hill, 1997.
[12] D. F. Specht, "A General Regression Neural Network", IEEE Transactions on Neural Networks, 2, (6), 1991, pp.568-576.