Gaining Insights into Road Traffic Data through Genetic Improvement

Anikó Ekárt, Alina Patelli, Victoria Lush
Aston Lab for Intelligent Collectives Engineering (ALICE)
Aston University
Birmingham, UK
{a.ekart,a.patelli2,v.lush1}@aston.ac.uk

Elisabeth Ilie-Zudor
Computer and Automation Research Institute
Hungarian Academy of Sciences
Budapest, Hungary
ilie@sztaki.mta.hu

ABSTRACT

We argue that Genetic Improvement can be successfully used for enhancing road traffic data mining. This would support the relevant decision makers with extending the existing network of devices that sense and control city traffic, with the end goal of improving vehicle flow and reducing the frequency of road accidents. Our position results from a set of preliminary observations emerging from the analysis of open access road traffic data collected in real time by the Birmingham City Council.

KEYWORDS

Genetic Improvement, symbolic regression, data mining

ACM Reference format:
Anikó Ekárt, Alina Patelli, Victoria Lush and Elisabeth Ilie-Zudor. 2017. Gaining Insights into Road Traffic Data through Genetic Improvement. In *Proceedings of GECCO '17 Companion, Berlin, Germany, July 15-19, 2017,* 2 pages.
DOI: http://dx.doi.org/10.1145/3067695.3082523

1 INTRODUCTION

The science of road traffic data - from eliminating irrelevant or corrupted records to actual data mining, such as classification and prediction - is essential to automated road traffic control [2, 3]. For example, extracting a model from a city's historical data about vehicle speed and incident rates would make it possible to predict the likelihood of a motor accident in a newly built junction. This insight could inform decisions about automated road traffic control, such as establishing the optimal speed limit, installing the ideal number of speed bumps or replacing the junction with a roundabout. This could ultimately cut costs, reduce pollution and save lives.

In spite of the notable progress made in terms of expanding the network of road traffic sensors, the data collected by those is challenging to analyse effectively. One typical problem would be predicting the occupancy of a newly built parking lot by combining existing models generated for similar locations, rather than waiting to collect data on the new structure. We make the case that Genetic Improvement (GI) is a suitable tool to address such issues. GI is a search technique that specialises and combines existing software to produce new and enhanced algorithms [4, 7], as opposed to Genetic Programming (GP) that traditionally starts either from scratch [8] or a heuristically extracted primitive set [6].

We base the argument on two pivotal points:

- the documented success of symbolic regression in terms of providing insight into large datasets [1] and
- the promising potential of GI stemming from its capacity to combine and exploit existing software rather than starting search from scratch (a feature called "code scavenging" [7]).

2 THE CHALLENGES OF MINING OPEN ACCESS ROAD TRAFFIC DATA

The 'Birmingham in real-time' project is supported by the Birmingham City Council and provides open source data[1] from various road traffic sensors, such as inductive loops, cameras feeding images into automatic number plate recognition (ANPR) algorithms, etc. The Birmingham city area is 103.4 mi^2 (approximately 11.5 miles in diameter) with over 400 road traffic sensors installed at various junctions in the city. The historic data available for the past two years is updated periodically (at five minute intervals) by collecting, collating and uploading new sensor readings. Output data include average speed and traffic flow, that is, the number of vehicles passing detected by the sensor, within the five minutes time frame, as well as travel time between two nodes of the observed grid.

Two important challenges of working with large repositories of data collected by physical (thus fallible) devices are accuracy and completeness [2]. Road traffic data is no exception - we note the following key issues often recurring in the Birmingham case study:

- **Incorrect data.** Sensors exposed to the elements are affected by significant brightness and temperature variations, causing inaccuracies in the data they collect. Those values are difficult to tell apart from correct ones, e.g., a traffic flow sensor that went offline and one that monitors an area where no car has passed during one sampling interval, will both record a value of 0. Erroneous readings may also originate from incorrectly installed sensors (using one inductive loop to monitor multi-lane junctions is a common infrastructure flaw).
- **Unavailable data.** Some features, such as the geolocation description of vehicle flow detectors, are characterised by a significant proportion of missing values (52% in the case of the data from the 21st of March 2017).

[1] http://butc.opendata.onl/AL_OpenData

GECCO '17 Companion, Berlin, Germany

978-1-4503-4939-0/17/07…$15.00
DOI: http://dx.doi.org/10.1145/3067695.3082523

We consider the following research questions:

(1) **Identify good locations for new sensors.** The decision making process, when it comes to extending busy roads or setting up intermediate junctions on an existing road, should be informed by insights extracted from available data - a goal difficult to achieve when the reliability of that data is questionable.
(2) **Discover subtle relationships between data streams.** It is sensible to expect a strong correlation between data sets recorded by two traffic flow sensors in consecutive junctions on a busy road. However, it would be even more useful to automatically detect understated, unexpected dependencies between different sensors' output.
(3) **Pre-process sensor data.** The quality of data mining generated models relies heavily on the accuracy and completeness of the input data. Automatically identifying flawed records (transmitted by faulty sensors) and differentiating between those and stand-by values (transmitted by working sensors recording no traffic flow) would be beneficial.
(4) **Predict useful features for existing sensors.** The cost of infrastructure maintenance may be reduced by having access to reliable estimates of road traffic density, at rush hour, next to a school or of midday occupancy in a city centre car park, etc. Such estimates may be difficult to obtain with substantial missing data.

3 GI FOR ROAD TRAFFIC PREDICTION

For the purpose of this position paper, we focus on the first practical question of where to add new sensor nodes in a city, when extending the road traffic sensing network, (1) above, which also involves locating dependencies between sensors (2).

When a partial network of sensors with data collected over a substantial period of time (i.e. two years in our case) exists, symbolic regression can be used very effectively to produce robust predictors of road traffic flow at every node in the partial network, but various time series models can also be employed. These predictors, irrespective of what method was used to derive them, can then form the starting code base for GI. As postulated by White and Singer [7], we are proposing to employ GI by using existing code, i.e., symbolic regressors and time series models created for existing nodes in the network, as ready-made functionality for creating models for sensor nodes with similar characteristics and new sensor nodes in the network. Mutation and crossover will be used to identify the modifications and combinations of existing models that lead to robust alternative models for such similar nodes. In the case of a new sensor node, it would be impossible to immediately model traffic flow as actual data collected at that location usually does not exist yet. One would need to collect data over a period of time first and then create the traffic flow model for this sensor based on the collected data. Therefore, if we can establish dependencies between sensors and similarities between groups of sensors, then, by using GI, appropriate combinations of existing road traffic flow models can be found to predict what the new sensor is expected to collect.

Figure 1 illustrates a possible scenario using the so-called primal graph representation for transport networks [5], where edges and vertices in the graph correspond to links (roads) and nodes (junctions) in the transport network. The part of the network including nodes **A-F** has similar layout to the part of the network including nodes **a-f** and this is indicated by the pairwise correspondence of nodes **A-a, B-b, C-c, D-d, E-e, F-f**. Consider that all these nodes except for **a** are fully equipped with sensors and the introduction of sensors at node **a** is being considered. Starting from the traffic flow models for the part of the network **A-F**, GI can be used to find the modifications (mutations) of these models to establish and validate models for nodes **b-f** and predict models for node **a**, where the new sensor insertion is being considered.

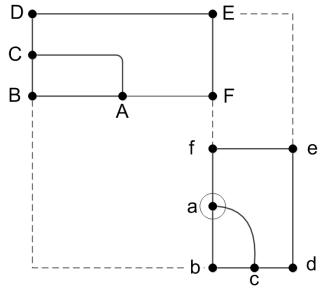

Figure 1: New road traffic sensor introduction: an example

As illustrated above, what-if scenarios can be provided to road traffic data collection decision makers to enable them to make informed decisions on the placement of the new sensors.

4 CONCLUSION

We argue that applying genetic improvement to incorporate existing code into automated programming, i.e., "code scavenging", is an avenue worth following for the problem of road traffic flow prediction at new sensor locations, based on established models for road traffic flow at pre-existing sensor locations. This is a complex real life problem on which the capability of GI can be demonstrated.

5 ACKNOWLEDGMENT

This work was supported by the European Commission through the H2020 project EXCELL (http://excell-project.eu/), grant No. 691829.

REFERENCES

[1] Michael Affenzeller, Stephan M Winkler, Gabriel Kronberger, Michael Kommenda, Bogdan Burlacu, and Stefan Wagner. 2014. Gaining deeper insights in symbolic regression. In *Genetic Programming Theory and Practice XI*. Springer, 175–190.
[2] Andrew Hamilton, Ben Waterson, Tom Cherrett, Andrew Robinson, and Ian Snell. 2013. The evolution of urban traffic control: changing policy and technology. *Transportation planning and technology* 36, 1 (2013), 24–43.
[3] Falilat Jimoh and Thomas Leo McCluskey. 2016. Self-management in urban traffic control–an automated planning perspective. Springer.
[4] William B. Langdon and Mark Harman. 2012. Genetically improving 50000 lines of C++. Research Note RN/12/09. (2012).
[5] Stephen Marshall. 2016. Line structure representation for road network analysis. *The Journal of Transport and Land Use* 9, 1 (2016), 29–64.
[6] Riccardo Poli, William B Langdon, Nicholas F McPhee, and John R Koza. 2008. *A field guide to genetic programming*. Lulu. com.
[7] David R White and Jeremy Singer. 2015. Rethinking genetic improvement programming. In *Proceedings of the Companion Publication of the 2015 Annual Conference on Genetic and Evolutionary Computation*. ACM, 845–846.
[8] John R Woodward, Colin G Johnson, and Alexander EI Brownlee. 2016. GP vs GI: if you can't beat them, join them. In *Proceedings of the 2016 on Genetic and Evolutionary Computation Conference Companion*. ACM, 1155–1156.