A Bus Delay Prediction Model Based on Machine Learning Algorithms

Carl Ka To Ma, Hugo Fang, Cato Yuan Wen
Shanghai High School International Division

***Abstract*—Bus delay prediction is a practical feature made available in many smart cities, allowing people to estimate their travel time when taking public transportation. Existing algorithms are built on the prerequisite that every vehicle in the bus network is equipped with GPS capable of sharing real-time location. This is not the case in developing cities. Therefore, the aim of this paper is to introduce a bus delay prediction model that only requires common, accessible data like date and time, weather, bus route, and traffic incident reports. While these broad categories of data have little meaning, they are preprocessed into specific factors that show a much stronger correlation with bus delay, such as peak or off-peak hours, visibility, precipitation, etc. A total of 24 input features are passed into a feed-forward neural network, trained with data collected in Toronto, Canada, over the past six years. While Toronto is used as an example because it provides convenient access to open data, the model is designed to work in all cities around the world. The performance of this algorithm is satisfactory: its 81% accuracy, despite being far from the near-100% accuracy archived by algorithms utilizing real-time GPS location, is fairly acceptable in a developing city before advanced systems are installed.**

***Index Terms*—Bus delay, delay prediction, machine learning, neural network, open data, smart cities**

I. Introduction

The emergence of smart cities fueled by advances in Artificial Intelligence (AI) changed the way people live over the past decade. From optical character recognition to automated driving systems, AI has brought countless benefits, simplifying human tasks or even achieving ones otherwise impossible by humans. One prerequisite of AI, however, is large sets of training data. This problem is hopefully solved in smart cities, where data is collected from the infinitely many sensors belonging to the Internet of Things and utilized to improve the quality of city services. Furthermore, according to the European Commission, a smart city is one with an efficient transport network, an upgraded water supply, a more interactive and responsive city administration, and so on. The focus of this essay will be on the first characteristic—transportation.

A. Background

Transportation has been revolutionized in smart cities with the introduction of applications like Google Maps, which provide a variety of services including route planning, real-time traffic information, and more. These services are made possible by the joint efforts of the mapping service provider (e.g. Google Maps) and the local transport department, responsible for the technology and the access to data, respectively. In some developed cities, the two offices are able to collaborate successfully, enhancing their services and bringing added convenience to the public. One example of a successful implementation is in the city of Hong Kong (HKSAR). Not only does Google Maps return a feasible route when the user enters a starting point and destination, it also considers factors like the bus schedule and potential delays if one chooses to take public transportation. According to Fabrikant (2019) from Google Research, bus delay predictions adopted by Google Maps rely on a sophisticated machine learning model of which details are not disclosed. As a result, an additional literature review is conducted to examine existing pieces of research and methodologies on the implementation of bus delay predictions.

B. Literature Review

The significance of this topic and the potential benefits it brings urge researchers worldwide to conduct similar explorations. In terms of algorithms, early researchers like Angelo (1999) and Patnaik (2004) used statistical models—namely, nonlinear time series and multivariable regression. Each of these early algorithms has their own limitations when applied to the prediction of bus delays, as Yin, Zhong, et al. (2006) have summarized: low accuracy, poor performance, or invalid assumptions. Yin and Zhong then suggested that the rapid development of machine learning enables the modeling of bus delays with support vector machines (SVM) and artificial neural networks (ANN). Indeed, these machine learning algorithms are becoming the spotlight of discussion. Sjafjell, Dahl, and Skogen (2014) proposed means to optimize the SVM and ANN models, eventually reaching a conclusion that their implementation of the SVM model would outperform the ANN model with less training, while the opposite holds true given more training. In terms of input parameters, Patnaik (2004) collected data specifically for his multivariable regression model, including the exact number of passengers boarding or exiting the bus. Recent models featuring machine learning are capable of processing large datasets, so researchers are inspired to use the real-time location of buses as an input parameter, persuading cities to install GPS systems on every vehicle. Sjafjell's SVM and ANN models, for example, analyze the data collected by the GPS trackers during the three previous trips.

C. Objectives

Utilizing real-time GPS location as an input parameter in a bus delay prediction model, as Google Maps must have done