Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 176 (2020) 3009–3018

Procedia Computer Science

www.elsevier.com/locate/procedia

24th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems

Predictive analytics on open big data for supporting smart transportation services

Paul Patrick F. Balbin[a], Jackson C.R. Barker[a], Carson K. Leung*[a], Marvin Tran[a], Riley P. Wall[a], Alfredo Cuzzocrea[b]

[a] *University of Manitoba, Winnipeg, MB, Canada*
[b] *University of Calabria, Rende, Italy*

Abstract

In the current era of big data, huge quantities of valuable data, which may be of different levels of veracity, are being generated at a rapid rate. Embedded into these big data are implicit, previously unknown and potentially useful information and valuable knowledge that can be discovered by data science solutions, which apply techniques like data mining. There has been a trend that more and more collections of these big data have been made openly available in science, government and non-profit organizations so that people could collaboratively study and analysis these open big data. In this article, we focus on open big data for public transit because public transit (e.g., bus) as a means of transportation is a vital part of many people's lives. As time is a precious resource, bus delays could negatively affect commuters' plans. Unfortunately, they are inevitable. Hence, many existing works focused on predicting bus delays. However, predicting on-time or early buses is also important. For instance, commuters who come to a bus stop on time may still miss their buses if the buses leave early. So, in this article, we examine open big data about bus performance (e.g., early, on-time, and late stops). We analyze the data with frequent pattern mining and make predictions with decision-tree based classification. For illustration, we perform predictive analytics on real-life open big data available on Winnipeg Open Data Portal, about bus performance from Winnipeg Transit. It shows the benefits of predictive analytics on open big data for supporting smart transportation services.

Keywords: Predictive analytics, open data, Winnipeg open data, big data, transportation data, on-time performance, frequent patterns, software engineering and large-scale systems

1. Introduction

In today's technology based landscape, huge quantities of data are being generated at a rapid rate. Some data are readily available such as the case with meteorological data [6], sports results [28] and stock prices [31]. Others data require collection such as the case with web and social networks [7, 17, 29, 39]. Embedded into these *big data* [13, 23, 35, 37, 41, 42, 43] are implicit, previously unknown and potentially useful information and valuable knowledge that can be discovered by data science or data mining [21, 22]. Insights gained from the discovered

E-mail address: kleung@cs.umanitoba.ca

1877-0509
10.1016/j.procs.2020.09.202

knowledge can have benefits for scientific reasons such as classifying and segmenting data, as well as hypothesis formation. Other insights can be used for commercial reasons, such as providing better services.

Open data [16, 19] have become more popular. There has been a trend that more and more collections of these big data have been made openly available in science, government and non-profit organizations. Benefits of these open data include:

- added value such as (a) transparency allowing monitoring of government activities and (b) innovation allowing individuals opportunities to develop new applications, etc.; as well as
- gained insights and preservation to ensure information is saved and can be accessed in future times.

Consequently, many countries have contributed to the idea of *open government* by making government more accessible to everyone. According to the Organisation for Economic Co-operation and Development (OECD)[1], many of its member countries (e.g., South Korea, France, Ireland, Japan, Canada) have put efforts to

- make their public sector data available and accessible, as well as to
- support data re-use.

In Canada, many cities and provinces have joined the initiative towards providing open data acknowledging the value that is gained in doing so. According to the Canada Open Government Working Group (COGWG)[2], as of May 2020, there were 61 open municipalities (which include Vancouver[3], Winnipeg[4], Toronto[5], Ottawa[6], Montréal[7]), 12 open initiatives (e.g., Nunavut Geoscience[8]), and 12 "open provinces" (or more precisely, 10 open provinces including Manitoba[9] and two open territories) across Canada in providing open data.

One type of available data sets in these open data portals relates to *transit open data*. Many cities provide information about their transit systems, such as bus stop numbers, route numbers, etc. Specifically, for this article, transit on-time performance is of particular interest. Information such as bus delay times are provided by a few cities, but even fewer cities provide information about a bus arriving early. Among one of the cities that provide bus delay and bus earliness is the city of Winnipeg, where bus deviation from the scheduled time is openly available. This deviation is measured in seconds deviating away (both early and late) from the scheduled departure time.

To elaborate, every bus operated by Winnipeg Transit (a public transit agency in the City of Winnipeg) is equipped with an on-board computer and global positioning system (GPS), which logs the on-time performance of the bus as it leaves a bus stop. For example, if a bus was supposed to leave a bus stop at 10:00:00 but actually left at 10:03:23, then the information that the bus was 3 minutes and 23 seconds late leaving the bus stop (i.e., −203 seconds from its scheduled departure time) was logged by the on-board computer on the bus. As another example, if a bus was supposed to leave a bus stop at 10:00:00 but actually left at 09:58:23, then the information that the bus was 1 minute and 37 seconds early leaving the bus stop (i.e., +97 seconds) was logged by the on-board computer on the bus. The logged data are then collected and downloaded from the buses every day, and updated to the open data portal[10] on a monthly basis. Normally (e.g., prior to the reduction in weekday service due to the coronavirus disease 2019 (COVID-19) pandemic), Winnipeg Transit[11] operated a fleet of 640 buses over 93 bus routes departing (early, on-time, or late) from approximately 5,170 bus stops. As these buses operate for long hours (e.g., over 1.55M hours in

[1] http://www.oecd.org/gov/digital-government/open-government-data.htm
[2] https://open.canada.ca/en/maps/open-data-canada
[3] http://vancouver.ca/your-government/open-data-catalogue.aspx
[4] https://data.winnipeg.ca/
[5] http://www.toronto.ca/open
[6] http://ottawa.ca/en/city-hall/get-know-your-city/open-data
[7] http://donnees.ville.montreal.qc.ca/
[8] http://cngo.ca/
[9] http://www.manitoba.ca/openmb/
[10] https://winnipegtransit.com/en/open-data/on-time-performance/
[11] https://winnipegtransit.com/en/about-us/interestingtransitfacts

year 2018), the resulting open big data can be characterized by the well-known V's (e.g., 3 V's, . . ., 7 V's, . . ., 42 V's, 51 V's):

- Huge *volumes* of bus on-time performance data are being logged and collected (e.g., more than 90M logged instances in 2018).
- Useful information and knowledge embedded in these open big data are of high *value* and can be mined or discovered by techniques like data science or data mining.
- These open big data are logged and collected at a relatively high *velocity* (i.e., whenever a bus leaves a bus stop for every operating bus).
- Due to transmission problems, data issues, as well as GPS and other errors, these open big data may be of a different level of *veracity* (e.g., logged latitudes and longitudes of locations may not be too precise).

In general, on-time performance can be affected by numerous factors (e.g., construction, emergencies, special events, traffic congestion, weather events). According to Winnipeg Transit, on-time performance is defined as follows:

- a bus that left the stop more than 1 minute earlier than the scheduled departure time is considered *early*,
- a bus that left the stop within 1 minute before and 3 minutes after the scheduled departure time is considered *on time*, and
- a bus that left the stop more than 3 minutes later than the scheduled departure time is considered *late*.

There are reasons why predicting bus delay is important. In modern society, people rely on transit systems for an assortment of tasks. Going to work, attending sports games, meeting friends, getting to school and making it home are among the numerous possibilities for using transit systems. While frequent bus riders may rely on transit more than others, infrequent riders should be able to rely on transit as well. Traditionally, analyses on transit delay [3] have primarily drawn attention as the negative downsides of arriving late to a destination are quite high. For instance, employers or teachers do not appreciate tardiness. Employees arriving late may get reduced paid or lose their jobs. Students arriving late for a test may impact their grades. Similarly, individuals do not want to miss the beginning of a sports game or concert.

Often neglected is the impact of a bus leaving early, but it is also important to consider as well. The act of transferring and leaving the current bus to board another bus is a common occurrence as different routes will lead to different destinations. If the second bus departs early, individuals may miss the transfer and end up waiting for the next bus far longer than anticipated. They may have missed the last bus of the day because it left early. A different route could have been planned to avoid this. So, accurate predictions of bus earliness is also important.

Given the availability of both bus earliness and delay in Winnipeg open big data, it is logical to examine them. We wonder whether it is feasible to analyze the data for discovering frequently occurring patterns. We also wonder whether it is feasible to predict future bus earliness or delay based on historical data. To answer these questions, we examine open big data on bus performance—namely, both early and late stops in this article. We also analyze the data with frequent pattern mining and make predictions with decision-tree based classification. For illustration, we perform predictive analytics on real-life Winnipeg open big data about on-time bus performance. Our *key contributions* of this article include our predictive analytics on open big data for supporting smart transportation services.

The remainder of this article is organized as follows. The next section discusses related works. Section 3 presents our data mining for on open big data about on-time bus performance to discover frequent patterns and our decision-tree based classification for predicting early, on-time and late bus. Evaluation results and conclusions are provided in Sections 4 and 5, respectively.

2. Related works

Data science solutions have been applied to transportation data. For example, we [25, 26] previously proposed urban data mining algorithms to accurately classify ground transportation modes (e.g., bus, car, bike, or walk) of commuters by analyzing their GPS data, accelerometer data, and/or dwell time history.

Besides ground transportation mode classification, there have been works on using transportation data to predict travel time. To elaborate, in KES 2009, two algorithms—namely, Successive Moving Average (SMA) and Chain Average (CA) algorithms [5]—were developed, based on the concept of moving average, to predict travel time. In KES 2010, the Modified k-Means Clustering (MKC) algorithm [32] was designed to make more accurate prediction on travel time when compared with those predicted by the SMA and CA algorithms. In KES 2011, we [4] proposed the Improved Successive Moving Average (iSMA) and Improved Chain Average (iCA) algorithms, which use non-recursive equations to directly compute the predicted travel time. Both iSMA and iCA algorithms were shown to improve the travel time prediction process (through the reduction of both time and space requirements and the maintenance of prediction accuracy) when compared with the SMA and CA algorithms. Moreover, we [4] also proposed the Improved Modified k-Means Clustering (iMKC) algorithm to lower the dimension of clustering space and measure dissimilarity in a single dimension of travel times. The iMKC algorithm was shown to improve over the MKC algorithm in predicting travel time, and was designed for supporting intelligent transportation systems.

Predictive analytics have been conducted on various real-life situations outside the scope of transportation [1, 12, 15, 18, 20, 33]. For those works that dealt with transportation data, many focused on predicting bus arrival times. For instance, Sun et al. [38] predicted the bus arrival time based on a geographic information system (GIS)-based map-matching algorithm. Lin et al. [30], on the other hand, employed GPS data and automatic fare collection (AFC) system data to predict bus arrival times through the use of artificial neural networks (ANN). However, auxiliary information like GIS and/or AFC system data may not be easily accessible to the general public. Instead, Rajput et al. [34] examined New York City open data, which are accessible to the general public. By applying clustering to the New York open data, they (a) identified areas with high congestion and insufficient bus stops and (b) recommended the installation of new bus stops in those areas.

The most closely related work is our previous work [3] on an intelligent system for predicting the severity of bus delay. The system examines the bus delay data—which capture only late buses but *no* on-time or early buses—from the Toronto Transit Commission (TTC, which is a public transit agency in the City of Toronto). Due to the nature of the data, the system analyzes the delay-causing incidents and their relationships to bus route, time, day, location and direction for predicting the severity of bus delay. In contrast, in the current article, we examine the bus performance data, which capture data about *on-time and early buses* in addition to late buses.

3. Predictive analytics on open big data

One of the biggest problems facing public transportation, at least as far as consumers are concerned, is reliability. So, designing and developing a reliable method to predict whether or not a specific bus would arrive early, on time, or late is in demand. Such a method would help provide information to commuters for making better plans for their commutes. In this section, we examine open big data on bus performance (e.g., past data regarding buses, their routes, day and time) with an aim to (a) discover frequently occurring patterns about bus performance and to (b) make prediction on the bus performance (e.g., early, on time, and/or late) for future bus based on past instances.

3.1. Data Preprocessing

To examine bus performance, we examine open big data from Winnipeg Open Data Portal and Winnipeg Transit. The data come in two collections:

- detailed on-time performance data
- aggregated on-time performance data by route and day

Between the two, the detailed on-time performance data consist of the following nine attributes:

1. row ID, which is a unique identifier in the dataset;
2. stop number, which is a unique 5-digit number assigned to an individual bus stop;
3. route number, which captures a unique 1- to 3-digit bus route number;
4. route name, which captures the common name that corresponds to a route number;

5. route destination, which indicates the terminus of a route and helps determine the direction (e.g., northbound vs. southbound) of the bus;
6. day type, which takes on one of the three values:

    - Sunday,
    - weekday (i.e., Monday to Friday), or
    - Saturday;

7. scheduled time, which is a timestamp indicating the departure date and time;
8. deviation, which captures the deviation (in seconds) between the scheduled departure time and the actual departure time from a stop:

    - a negative deviation indicating a bus departs from a stop later than its scheduled time, whereas
    - a positive deviation indicating a bus departs from a stop earlier than its scheduled time; and

9. location, which is a point—i.e., GPS coordinates in the form of POINT (*longitude*, *latitude*)—representing the location of a particular bus stop.

Based on these nine attributes in the detailed on-time performance data, we observed the following:

- In general, buses run on a different schedule among weekdays, Saturdays, and Sundays. In particular, buses are the most frequent on weekdays, less frequent on Saturdays, and the least frequent on Sundays. The attribute "day type" helps distinguish data from one day type (e.g., weekday) from another (e.g., Saturday).
- Recall the definitions of early, on-time, and late bus from Section 1. An integer value lower than -180 for the attribute "deviation" implies that the bus departed more than 180 seconds later than its scheduled departure time, which counts as an instance of a *late stop*. In contrast, a "deviation" value between -180 and 60 inclusive implies that the bus departed within $[-180, 60]$ seconds from its scheduled departure time (i.e., no more than 60 seconds earlier and no more than 180 seconds later than the scheduled time), which counts as an *on-time stop*. Finally, a "deviation" value higher than 60 implies that the bus departed more than 60 seconds earlier than its scheduled departure time, which counts as an *early stop*.
- For the attribute "location", all bus stop points are represented as (*longitude*, *latitude*)-GPS coordinates in their decimal degrees (DD). Positive longitudes are east of the Prime Meridian, and negative ones are west. Positive latitudes are north of the Equator, and negative ones are south. As Winnipeg is located in the northern and western hemisphere, GPS-coordinates of stop location are expected to be of negative longitudes and positive latitudes. More specifically, they are expected to be around 49.8951°N, 97.1384°W. For example, Stop number 60674 located in the main bus loop at University of Manitoba is represented as POINT (−97.1326823838922 49.8077993749079), which indicates the geographical location of 49°48′28.1″N 97°07′57.5″W in degree-minute-second (DMS) format.

In addition to the detailed on-time performance data, the open big data also contain the aggregated on-time performance data. The aggregated data consist of the following nine attributes, in which the first four are the same as in the detailed data whereas the latter five are aggregated from attributes in the detailed data:

1. route number
2. route name
3. route destination
4. day type
5. day, which captures the departure date (but not time) from the scheduled time
6. time period, which captures the departure time (but not date) from the scheduled time and bins it into one of the five periods within a day:

- 05:00-09:00, which refers to morning peak hours;
- 09:00-16:00, which refers to off-peak hours;
- 16:00-18:30, which refers to afternoon peak hours;
- 18:30-22:30, which refers to evening hours;
- 22:30-05:00, which refers to night hours

7. total number of early stops, which indicates the number of occurrences when bus departs a stop more than 1 minute before its scheduled departure time
8. total number of late stops, which indicates the number of occurrences when bus departs a stop more than 3 minutes after its scheduled departure time
9. total number of on-time stops, which indicates the number of occurrences when bus departs a stop within 1 minute before and 3 minutes after its scheduled departure time

Between these two collections of bus on-time performance data, we examine the detailed on-time performance data because the aggregated one can be derived from the detailed one. Moreover, some fine-grained information is lost in the aggregated data. More interesting knowledge can be discovered from the detailed data.

Among the nine attributes in the detailed bus on-time performance data, we perform the following during our data preprocessing step:

- ignore the row ID for the analytics as it is just a unique identifier in the dataset
- capture the *stop number*
- capture the *route number*
- ignore the route name as there is a 1-to-1 correspondence to (or dependence on) the route number
- map the route destination into an integer value to indicate the *route direction*:
  - 1 to indicate destinations for southbound or westbound buses
  - 2 to indicate destinations for northbound or eastbound buses
  - other single-digit integer to indicate additional destinations (e.g., 3 to indicate destinations for southbound or westbound buses for Sundays)
- extract the *day of the week* from the timestamp for the scheduled time (instead of capturing the day type) because commuters' behavior may be different among the five weekday (e.g., due to students on Monday-Wednesday-Friday schedule vs. those on Tuesday-Thursday schedule, employees who work 4.5 days, people who left early on Fridays, holidays that fall on Mondays)
- extract the time (but not date) from the timestamp for the scheduled time and bins it into one of the five *time periods* within a day:
  - 05:00-09:00, which refers to morning peak hours;
  - 09:00-16:00, which refers to off-peak hours;
  - 16:00-18:30, which refers to afternoon peak hours;
  - 18:30-22:30, which refers to evening hours;
  - 22:30-05:00, which refers to night hours
- bin deviation (between the scheduled departure time and the actual departure time from a stop) into one of the seven *severity levels*:
  - Extremely late: deviation of < -1800 seconds, i.e., more than 30 minutes later than the scheduled time
  - Very late: deviation of $[-1800, -600)$ seconds, i.e., more than 10 minutes but no more than 30 minutes later than the scheduled time
  - Slightly late: deviation of $[-600, -180)$ seconds, i.e., more than 3 minutes but no more than 10 minutes later than the scheduled time
  - On-time: deviation of $[-180, 60]$ seconds, i.e., no more than 1 minute earlier and no more than 3 minutes later than the scheduled time
  - Slightly early: deviation of (60, 600] seconds, i.e., more than 1 minutes but no more than 10 minutes earlier than the scheduled time

- Very early: deviation of (600, 1200] seconds, i.e., more than 10 minutes but no more than 20 minutes earlier than the scheduled time
- Extremely early: deviation of > 1200 seconds, i.e., more than 20 minutes earlier than the scheduled time

This preprocessing step results in extracting six attributes for every data point for our analysis.

3.2. Frequent pattern mining

Once the collection of detailed on-time performance data was preprocessed, we then examine the resulting data. To facilitate our data analytic process, we map some categorical attributes into numerical values:

- For attribute "day of the week", we map Monday into 1, Tuesday into 2, . . ., Saturday into 6, and Sunday into 7.
- For attribute "time period", we map morning peak hours into 1, off-peak hours into 2, afternoon peak hours into 3, evening hours into 4, and night hours into 5.
- For attribute "severity level", we map extremely late into 1, very late into 2, slightly late into 3, on-time into 4, slightly early into 5, very early into 6, and extremely early into 7.

Then, we use data mining, machine learning, as well as statistical machine intelligent and learning engine (Smile), for our analytics. In particular, with frequent pattern mining, we discover frequently occurring patterns. The discovered frequent patterns help reveal bus on-time performance. To elaborate, the frequent patterns lead to new knowledge such as:

- For which routes, did the buses frequently depart the stops early, on-time or late?
- For which routes and directions, did the buses frequently depart the stops early, on-time or late?
- On which days of the week, did the buses frequently depart the stops early, on-time or late?
- At which time periods, did the buses frequently depart the stops early, on-time or late?
- What are relationships between routes and severity levels of early (on-time or late) bus?
- What are relationships between days of the weeks and severity levels of early (on-time or late) bus?
- What are relationships between time periods and severity levels of early (on-time or late) bus?
- What are frequent relationships among the six attributes?

3.3. Classification and prediction

Based on the discovered frequent patterns, association rules can be formed to reveal the association between attributes in the antecedent and those in the consequent of the rules. To a further extent, associative classification rules can be formed by putting the attribute "severity level" in the consequent and any of the other attributes in the antecedent of the classification rules. Consequently, these rules lead to new knowledge such as:

- Buses on which routes are likely to depart extremely/very/slightly late, on-time, or extremely/very/slightly early?
- Buses on which routes and directions are likely to depart the stops on-time (or at other severity levels)?
- On which days of the week, are buses likely to depart the stops on-time (or at other severity levels)?
- At which time periods, are buses likely to depart the stops on-time (or at other severity levels)?
- Under which combinations of the first five attributions would buses likely to depart the stops on-time (or at other severity levels)?

Besides associative classification, we also construct a decision tree by training the model with certain percentage of historical data and test the model with the remaining portion of historical data. The resulting decision-tree based classification helps identify the discriminatory attributes for explaining the classification model (i.e., decision tree).

4. Evaluation

Detailed on-time performance data from July 2017 to the end of the most recent month are available in the Winnipeg Open Data Portal. We downloaded the monthly detailed on-time performance data from July 2017 onwards. We apply frequent pattern mining to all available data, as well as to 12-month data (May 2019-April 2020). For classification and prediction, we use data from July 2017-June 2019 (i.e., oldest 24 months) for training and data from July-December 2019 (i.e., 6 months) for testing. This give a training-testing ratio of 80%:20%.

Evaluation was run on an Intel Core i7-4790 processor with 16GB of RAM. An average runtime was computed from the average of 5 runs. For instance, frequent pattern mining took about 204 seconds, whereas classification and prediction took about 228 seconds.

The predictive analytics on these open big data led to some interesting knowledge in the forms of frequent patterns, association rules, associative classification rules, decision trees, and predictions. Sample observations on this discovered knowledge include:

- 53% of buses operating on Route 181 departed bus stops (extremely/very/slightly) late, with a frequency of 113,845 late stops (i.e., late departures from the bus stops).
- 79% of buses operating on Route 642 departed bus stops on-time, with a frequency of 27,283 on-time stops.
- 79% of buses operating on Route 23 departed bus stops (extremely/very/slightly) early, with a frequency of 11,940 early stops.
- 77% of buses operating on October 11, 2019, departed bus stops (extremely/very/slightly) late, with a frequency of 197,825 late stops. Cross checking with the weather information, an explanation to such a high number of late stops was probably due to the inclement weather conditions (caused by an early blast of winter weather) on that day.
- 32% of buses operating during March 23-26, 2020, departed bus stops (extremely/very/slightly) early, with a frequency of 423,224 early stops over 4 days. Cross checking with the news, an explanation to such a high number of early stops was probably due to a sudden drop in ridership in these first four working days (when people started working from home) after the state of emergency was declared (in response to COVID-19 pandemic) on March 20.

In terms of prediction accuracy, initially, the classifier correctly predicted the severity levels of stops (e.g., extremely/very/slightly early or late, or on-time) on 29,337,870 instances among all 43,195,136 predictions, leading to a *precision* of around 68%. We realized the veracity of data (which was partially caused by transmission problems, data issues, as well as GPS and other errors), and those extremely early stops (i.e., bus departing more than 20 minutes earlier than its scheduled time) could be unreliable and likely to be noise. When training the classification model with the noise would lead to overfitting, which would negatively impact the model and its prediction accuracy. Hence, we removed the noise and train the model with random samples taken from some subset of the training data. Consequently, the precision was improved to around 71%. Further improvements are expected when fine-tuning some parameters. In addition to precision, we also measure the differences between the predicted severity level and the actual severity level of bus on-time performance in terms of *root-mean-square error* (*RMSE*):

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(\widehat{y_i} - y_i)^2}{n}} \tag{1}$$

where $\widehat{y_i}$ is a predicted severity level of bus on-time performance (e.g., extremely/very/slightly early/late, or on-time), y_i is an actual severity level of bus on-time performance, and n is the total number of stop instances. The results show that the RMSE of the initial classifier was 0.880863, which was reduced to 0.741962 for the improved classifier.

5. Conclusions

With the high availability of open data, data mining has become a very useful way to unlock secrets and find value in data that would not have been easily discovered otherwise. The City of Winnipeg provides an open big data about on-time performance for Winnipeg Transit through its open data portal. Unlike similar data that focus solely on bus delay,

these Winnipeg data provide a broader coverage as they cover late buses, as well as early and on-time buses. Given the availability of these open big data, whether it is feasible to analyze the data for discovering frequently occurring patterns? Whether it is feasible to predict future bus earliness or delay based on historical data? We answered these questions in this article. Specifically, we examined and analyzed the open big data on bus performance with frequent pattern mining, and made predictions with decision-tree based classification. When performing predictive analytics on these real-life Winnipeg, we discovered interesting knowledge in the forms of frequent patterns, association rules, associative classification rules, decision trees, and predictions. Frequent patterns reveal frequently occurring attribute-values. Association rules reveal a high likelihood of having some particular attribute-values in the consequent of the rules given some other attribute-values in the antecedent of the rules. Associative classification rules and decision trees help predicting any of the seven severity levels of bus on-time performance. All these show the feasibility of analyzing big open data to discover frequent patterns and to predict future bus on-time performance based on the historical data. This article also shows the benefits of predictive analytics on open big data for supporting smart transportation services.

As ongoing and future work, we explore different directions to further enhance our predictive analytics on open big data. For instance, we examine graph representation and analytics [2, 11, 36, 40] of transportation data. We investigate the use of online analytical processing (OLAP) [8, 9, 10] for multi-dimensional analysis (MDA). We exploit deep learning [14, 24] in predicting the severity levels of bus on-time performance. We also incorporate data from other sources (e.g., weather information, news) for explanations of our observations of the discovered knowledge (e.g., why buses were very late on a certain day?). Moreover, based on the knowledge learned from the on-time performance about buses in Winnipeg, we would like to apply transfer learning [27] to other municipalities and other transportation modes such as (a) bus, streetcar and subway in Toronto, as well as (b) bus and métro in Montréal.

Acknowledgements

This project is partially supported by the Natural Sciences and Engineering Research Council of Canada (NSERC), as well as the University of Manitoba.

References

[1] Ang, R.P., & D.H. Goh (2013) “Predicting juvenile offending: a comparison of data mining methods,” Int. J. Offender Therapy and Comparative Criminology 57(2), pp. 191-207.
[2] Ashraf, N., R.R. Haque, M.A. Islam, C.F. Ahmed, C.K. Leung, J.J. Mai, & B.H. Wodi (2019) “WeFreS: weighted frequent subgraph mining in a single large graph,” in ICDM 2019, pp. 201-215.
[3] Audu, A.A., A. Cuzzocrea, C.K. Leung, K.A. MacLeod, N.I. Ohin, & N.C. Pulgar-Vidal (2019) “An intelligent predictive analytics system for transportation analytics on open data towards the development of a smart city,” in CISIS 2019, pp. 224-236.
[4] Chowdhury, N.K., & C.K. Leung (2011) “Improved travel time prediction algorithms for intelligent transportation systems,” in KES 2011, Part II, pp. 355-365.
[5] Chowdhury, N.K., R.P.D. Nath, H. Lee, & J. Chang (2009) “Development of an effective travel time prediction method using modified moving average approach,” in KES 2009, Part I, pp. 130-138.
[6] Cox, T.S., C.S.H. Hoi, C.K. Leung, & C.R. Marofke (2018) “An accurate model for hurricane trajectory prediction,” in IEEE COMPSAC 2018, vol. 2, pp. 534-539.
[7] Cuzzocrea, A. (2006) “Combining multidimensional user models and knowledge representation and management techniques for making web services knowledge-aware”. WIAS 4(3), pp. 289-312.
[8] Cuzzocrea, A., & E. Bertino (2011) “Privacy preserving OLAP over distributed XML data: a theoretically-sound secure-multiparty-computation approach,” JCSS 77(6), pp. 965-987.
[9] Cuzzocrea, A., C. de Maio, G. Fenza, V. Loia, & M. Parente (2016) “OLAP analysis of multidimensional tweet streams for supporting advanced analytics,” in ACM SAC 2016, pp. 992-999.
[10] Cuzzocrea, A., & V. Russo (2009) “Privacy preserving OLAP and OLAP security,” in Encyclopedia of Data Warehousing and Mining, pp. 1575-1581.
[11] Cuzzocrea, A., & I. Song (2014) “Big graph analytics: the state of the art and future research agenda,” in DOLAP 2014, pp. 99-101.
[12] Czibula, G., A. Mihai, & L.M. Crivei (2018) “S PRAR: a novel relational association rule mining classification model applied for academic performance prediction,” Procedia Computer Science 159. KES 2019, pp. 20-29.
[13] Dedić, N., & C. Stanier (2016) “Towards differentiating business intelligence, big data, data analytics and knowledge discovery,” in ERP Future 2016, pp. 114-122.

[14] de Guia, J., M. Devaraj, & C.K. Leung (2019) "DeepGx: deep learning using gene expression for cancer classification," in IEEE/ACM ASONAM 2019, pp. 913-920.
[15] Friesen, J., L. Rausch, P. Pelz, & J. Fürnkranz (2018) "Determining factors for slum growth with predictive data mining methods," Urban Science 2(3), 81:1-81:19.
[16] Ivancevic, V., & I. Lukovic (2018) "National university rankings based on open data: a case study from Serbia," Procedia Computer Science 126. KES 2018, pp. 1516-1525.
[17] Jiang, F., C.K. Leung, R. Middleton, & A.G.M. Pazdor (2018) "Big social data mining in a cloud computing environment," in ICCBB 2018, pp. 58-65.
[18] Jodayree, M., M. Abaza, & Q. Tan (2019) "A predictive workload balancing algorithm in cloud services," Procedia Computer Science 159. KES 2019, pp. 902-912.
[19] Kassen, M. (2013) "A promising phenomenon of open data: a case study of the Chicago open data project," Government Information Quarterly 30(4), pp. 508-513.
[20] Kulla, E., S. Morita, K. Katayama, & L. Barolli (2018) "Route lifetime prediction method in VANET by using AODV routing protocol (AODV-LP)," in CISIS 2018, pp. 3-11.
[21] Lakshmanan, L.V.S., C.K. Leung, & R.T. Ng (2000) "The segment support map: scalable mining of frequent itemsets," ACM SIGKDD Explorations 2(2), pp. 21-27.
[22] Leung, C.K. (2009) "Frequent itemset mining with constraints," in Encyclopedia of Database Systems, pp. 1179-1183.
[23] Leung, C.K. (2018) "Big data analysis and mining," in Encyclopedia of Information Science and Technology, 4e, pp. 338-348.
[24] Leung, C.K., P. Braun, & A. Cuzzocrea (2019) "AI-based sensor information fusion for supporting deep supervised learning," Sensors 19(6), pp. 1345:1-1345:12.
[25] Leung, C.K., P. Braun, C.S.H. Hoi, J. Souza, & A. Cuzzocrea (2019) "Urban analytics of big transportation data for supporting smart cities," in DaWaK 2019, pp. 24-33.
[26] Leung, C.K., P. Braun, & A.G.M. Pazdor (2018) "Effective classification of ground transportation modes for urban data mining in smart cities," in DaWaK 2018, pp. 83-97.
[27] Leung, C.K., A. Cuzzocrea, J.J. Mai, D. Deng, & F. Jiang (2019) "Personalized DeepInf: enhanced social influence prediction with deep learning and transfer learning," in IEEE BigData 2019, pp. 2871-2880.
[28] Leung, C.K., & K.W. Joseph (2014) "Sports data mining: predicting results for the college football games," Procedia Computer Science 35. KES 2014, pp. 710-719.
[29] Leung, C.K., S.K. Tanbeer, & J.J. Cameron (2014) "Interactive discovery of influential friends from social networks," Social Network Analysis and Mining 4(1), pp. 154:1-154:13.
[30] Lin, Y., X. Yang, N. Zou, & L. Jia (2013) "Real-time bus arrival time prediction: case study for Jinan," China J. Transport. Eng. 139(11), pp. 1133-1140.
[31] Morris, K.J., S.D. Egan, J.L. Linsangan, C.K. Leung, A. Cuzzocrea, & C.S.H. Hoi (2018) "Token-based adaptive time-series prediction by ensembling linear and non-linear estimators: a machine learning approach for predictive analytics on big stock data," in IEEE ICMLA 2018, pp. 1486-1491.
[32] Nath, R.P.D., H. Lee, N.K. Chowdhury, & J. Chang (2010) "Modified k-means clustering for travel time prediction based on historical traffic data," in KES 2010, Part I, pp. 511-521.
[33] Phankokkruad, M., & S. Wacharawichanant (2018) "Prediction of mechanical properties of polymer materials using extreme gradient boosting on high molecular weight polymers," in CISIS 2018, pp. 375-385.
[34] Rajput, P., Toshniwal, D., Agggarwal, A.: Improving infrastructure for transportation sys-tems using clustering. In: BDA 2017. LNCS, vol. 10721, pp. 129-143 (2017)
[35] Sassi, M.S.H., F.G. Jedidi, & L.C. Fourati (2019) "A new architecture for cognitive internet of things and big data," Procedia Computer Science 159. KES 2019, pp. 534-543.
[36] Singh, S.P., C.K. Leung, F. Jiang, & A. Cuzzocrea (2019) "A theoretical approach to discover mutual friendships from social graph networks," in iiWAS 2019, pp. 212-221.
[37] Snijders, C., U. Matzat, & U. Reips (2012) "'Big data': big gaps of knowledge in the field of internet," Int. J. Internet Science 7, pp. 1-5.
[38] Sun, D., H. Luo, L, Fu, W. Liu, X. Liao, & M. Zhao (2007) "Predicting bus arrival time on the basis of global positioning system data," Transp. Res. Rec. 2034(1), 62-72 (2007)
[39] Tanbeer, S.K., C.K. Leung, & J.J. Cameron (2014) "Interactive mining of strong friends from social networks and its applications in e-commerce," JOCEC 24(2-3), pp. 157-173.
[40] Upoma, F.M., S.A. Khan, C.F. Ahmed, T. Alam, S.A. Zahin, & C.K. Leung (2019) "Discovering correlation in frequent subgraphs," in IMCOM 2019, pp. 1045-1062.
[41] Cuzzocrea, A., Mastroianni, C., & Grasso, G.M. (2016) "Private databases on the cloud: Models, issues and research perspectives," in BigData 2016, pp. 3656-3661.
[42] Camara, R.C., Cuzzocrea, A., Grasso, G.M., Leung, C.K., Powell, S.B., Souza, J., & Tang, B. (2018) "Fuzzy Logic-Based Data Analytics on Predicting the Effect of Hurricanes on the Stock Market," in FUZZ-IEEE 2018, pp. 1-8.
[43] Braun, P., Cuzzocrea, A., Leung, C.K., Pazdor, A.G.M., Tanbeer, S.K., & Grasso, G.M. (2018) "An Innovative Framework for Supporting Frequent Pattern Mining Problems in IoT Environments," in ICCSA 2018, pp. 642-657.