An Innovative Fuzzy Logic-Based Machine Learning Algorithm for Supporting Predictive Analytics on Big Transportation Data

Carson K. Leung[1(✉)], Jonathan D. Elias[2], Shael M. Minuk[2], A. Roy R. de Jesus[1], and Alfredo Cuzzocrea[3]
[1]*Department of Computer Science, University of Manitoba*, Winnipeg, MB Canada
kleung@cs.umanitoba.ca
[2]*Department of Electrical and Computer Engineering, University of Manitoba*, Winnipeg, MB Canada
[3]*iDEA Lab, University of Calabria*, Rende, Italy

***Abstract*—In the current era of high precision monitoring and big data, many public transit users are still suffering from problems caused by transit delays. To help address this problem, we design and develop an innovative fuzzy logic-based machine learning algorithm for supporting predictive analytics on big transportation data to helps detect and predict the delay of some modes of public transport. To demonstrate the usefulness of our machine learning algorithm as a solution to this problem, we use it on heterogeneous data—namely, transit data and weather data—to predict the expected delay of streetcars (aka trolley cars) in the Canadian city of Toronto. To make accurate prediction, our algorithm takes into account multiple factors such us rain, snow, temperature, time of day, and season. Evaluation results show the effectiveness and usefulness of our fuzzy logic based machine learning algorithm for predictive analytics on big transportation data, which is promising toward development of a predictive intelligent transport system (ITS).**

***Index Terms*—computational intelligence, fuzzy systems, temporal data analysis, time series analysis, prediction, forecasting, data mining, knowledge discovery, big data applications, Toronto, streetcar, delay, weather, frequent pattern mining, random forest, regression, intelligent transport system (ITS)**

I. Introduction

In the current era of big data, high volumes of valuable data have been generated and collected from a wide variety of rich data sources. Examples of these big data include disease reports, epidemic data and statistics [32], financial data from stock market [27], genomic data [9], [29], music records [24], shopper market basket transactions [2], [20], [26], social media data [6], [14], [22], [23], [31], [34], transit data [21], as well as and weather data. Embedded in these data is implicit, previously unknown and potentially useful information and knowledge that can be discovered—by machine learning and data mining algorithms [4]—for social good. For example, analyzing and mining disease reports and epidemic data (for outbreak or disease analytics) helps people understand and predict the spread and severity of infectious disease like coronavirus disease 2019 (COVID-19) [32]. As another example, analyzing and mining shopper market basket transactions (for market basket analysis) helps reveal customer behaviour, which in turn could improve shopping experience. As a third example, analyzing and mining transit data (for predictive analytics in transportation industry) helps predict delays of public transit, which in turn could improve transit rider experience and enable intelligent transport system (ITS).

Usually, in numerous real-life applications, their associated data are related and heterogeneous. For example, social distancing, quarantine, isolation, and pharmaceutical interventions may affect the spread and severity of infectious disease like COVID-19. Hence, having additional information about these measures or interventions could enhance prediction accuracy of the disease. Similarly, as weather and promotional campaign may affect consumers' shopping patterns (e.g., heatwave or discount offers may increase the sales of ice cream), having this additional information could enhance accuracy of the analysis. As a third example, weather may affect the punctuality of public transit (e.g., heavy pouring rain reduces driver visibility may delay streetcars). Hence, having weather data in addition to transit data could enhance the prediction accuracy of delays in public transit.

In this paper, we focus on predictive analytics on big transportation data. Specifically, we present a fuzzy logic-based machine learning algorithm to predict transit delays by incorporating knowledge discovered from a wide variety of heterogeneous data:

- transit data, and
- weather data.

We first categorize data based on fuzzy logic, which helps capture the "degree of truth" of different categories of various features in the data. Then, we mine frequent patterns from integrated transit and weather data to determine important features contributing to the transit delays. Afterwards, we apply supervised learning—in particular, random forest regression—to predict transit delays. To illustrate and evaluate our algorithm, we conduct a case study on transportation data from the Canadian city of Toronto:

- Toronto Transit Commission (TTC)'s delay data from the City of Toronto's Open Data Portal[1], which capture delays of different transit modes (e.g., bus, streetcar, subway) and ridership information from 2014 to current.

[1] https://open.toronto.ca/

978-1-7281-6932-3/20/$31.00 ©2020 IEEE

- weather data (e.g., presence of snow on the ground, temperature, occurrence of rain, etc.) from Weather Dashboard for Toronto[2], which capture weather information[3] from the Canadian federal department of Environment and Climate Change Canada[4].

This study helps us examine what contributes to the streetcar delays within Toronto, and whether the weather has an impact as well. Hence, our *key contribution of this paper* is our innovative fuzzy logic-based machine learning algorithm for supporting predictive analytics on big transportation data.

The remainder of this paper is organized as follows. Next section presents background and related works. In Section III, we describe our fuzzy logic-based machine learning algorithm for predictive analytics on big transportation data. In Section IV, we illustrate key steps of our algorithm on a case study for predicting streetcar delays in the Canadian city of Toronto. Section V shows evaluation results, and Section VI draws the conclusions.

II. Background and Related Works

In this section, we present background on fuzzy logic and related works on transportation data analysis.

A. Fuzzy Logic

Fuzzy logic [5] is a form of multi-valued logic in which the truth values of variables may be any real number between 0 and 1, i.e., [0, 1] $\subseteq$ R. In other words, it uses the concept of partial truth for uncertainty [15], where the truth value may range between completely true and completely false. In contrast, Boolean logic captures the truth values of variables, which may only be the integer values 0 or 1. In contrast to the commonly used Boolean logic operators (AND, OR, NOT), fuzzy logic operators compute:

- MIN(x, y) for conjunction of two fuzzy logic values x and y;
- MAX(x, y) for disjunction of two fuzzy logic values x and y; and
- $1 - x$ for negation of a fuzzy logic value x.

Fuzzy logic is operated based on fuzzy set [37], which is often defined as (a) a trapezoid-shaped curve, or (b) a sigmoid function like $S(x) = \frac{1}{1+e^{-x}}$.

When compared with the probabilistic-based approaches that normally use (subjective) probability to express the likelihood of the factors, the fuzzy logic-based approaches usually use the concept of fuzzy logic and fuzzy set membership to model uncertainty and vagueness by expressing how much an observation on the factors falls within a vaguely defined set.

B. Analytics on Transportation Data

Over the past decade, researchers have proposed several prediction algorithms and transportation data analysis [3], [18], [19], [28], [33]. Many of them have focused on transport and traffic flow analytics [8], [35], [36]. A majority of these related works describe sophisticated techniques, including additive model [18], artificial neural network (ANN) [25], automatic personality classification (APC) [30], mean absolute percent error (MAPE), support vector regression (SVR) [17], and vehicle ad-hoc networks (VANETs) [19]. Some related works rely on dynamic data from advanced public transport system (APTS) [35], automatic fare collection (AFC) [25], and global position system [33]. Moreover, some related works incorporate dynamic data into the aforementioned sophisticated techniques through the use of Kalman filters [8], [25], [30], [35]. However, dynamic data may not necessarily be always available. Sophisticated techniques may incur costly computations. In contrast, our presented solution:

- only requires historical data (i.e., no requiring real-time tracking),
- takes advantage of widely tested and validated data mining techniques,
- can be easily developed and widely applicable to various transport data, as well as
- maintains a good fusion of computational complexity and prediction accuracy.

III. Our Fuzzy Logic-Based Predictive Analytics Algorithm

In this section, we describe our fuzzy logic-based machine learning algorithm for predictive analytics on big transportation data. As an overview, our algorithm consists of three key steps.

A. Key Step 1—Data Pre-Processing with Fuzzy Logic-Based Categorization

Our algorithm first categorizes data based on fuzzy logic, which helps capture the "degree of truth" of different categories of various features in the data. Specifically, we categorize the data by grouping similar data into broader categories. The main goal of this key step is to turn quantitative input features into similar categorical groups. These categorical groups can then be used for frequent pattern mining, data visualization amongst a categorical group, and input features for a machine learning model. Recall from Section I that we gather and combine the following heterogeneous data from multiple sources:

- For the transit data, the primary groups include the time of day and the length of delay.
- For the weather data, the primary groups include the amount of snow, average hourly temperature, the amount of rain, visibility, and average wind speed.

Due to the subjective nature of some of these primary groups (e.g., very poor, poor, moderate, and good visibility), we incorporate the concept of fuzzy logic to capture their "degree of truth".

B. Key Step 2—Frequent Pattern Mining

Once the data are pre-processed with fuzzy-logic based categorization, our algorithm then mines frequent patterns

[2] https://toronto.weatherstats.ca/

[3] https://weather.gc.ca/

[4] https://www.canada.ca/en/environment-climate-change.html

from integrated transit and weather data to determine important features contributing to the transit delay. Specifically, we apply association rule mining and/or frequent pattern mining techniques to discover frequent patterns—in the form of combinations of factors contributing to transit delays—from these pre-processed transportation data. These frequent patterns, in turn, give insight into the data to see what factors are frequently associated with one another. A main goal of this key step is to determine which combination of factors contribute to the different delay categories and if these factors were expected. Another main goal of this key step is to form interesting association rules, which can be served for *associative classification*, in which the antecedents of these rules are collection of frequently occurring factors contributing to the types/severity of transit delays in the consequents of these rules.

C. Key Step 3—Transit Delay Prediction with Supervised Learning

After mining frequent patterns in the form of combinations of factors contributing to transit delays, our algorithm then applies supervised learning to predict transit delay. The main goal of this key step is to develop an accurate model to predict the transit delay. Among different prediction approaches, a decision tree approach is simple, but the resulting prediction accuracy may be an issue. On the other hand, while a deep learning approach (e.g., convolutional neural network) usually leads to higher prediction accuracy, the prediction results may not be easily explainable. This is a trade-off between prediction accuracy and explainability. We hypothesized that weather—especially during the winter—will have a clear and statistically significant effect on the delay. We also observed that the transportation data (i.e., transit data augmented with weather data) are non-linear. Taking into consideration the trade-off and observations, a random forest regression would be a logical choice due to its reliability in inherently ranking important features, its prediction accuracy, and its simplicity for explaining the results. The resulting prediction is expected to have massive implications for the city, as well as businesses. For example, cities could use the accurately predicted delays to allocate resources and inform riders. As another example, businesses (e.g., peer-to-peer ride-sharing and ride hailing services such as Uber or Lyft) could use the accurately predicted delays to get a clear view of the actual demand within an area for their services, and thus accordingly adjust the price estimate for trips.

IV. Illustration of Our Algorithm with a Case Study on Predicting Toronto Streetcar Delays

In the previous section, we described the three key steps of our fuzzy logic-based machine learning algorithm for predictive analytics on big transportation data. Let us illustrate these three key steps with a case study on predicting delays based on transportation data on streetcars (aka trolley cars), which are transit data from the Toronto Traffic Commission (TTC) augmented with weather information.

A. Key Step 1—Transportation Data Pre-Processing with Fuzzy Logic-Based Categorization

An example of our categorical pre-processing step can be observed in Table I. This process was repeated for all categorical weather features, as well as time-based input features from the TTC delay data. For the time of day (TOD), we split the current time into four categories as outlined in Table I.

Besides the TOD, we also used many features from the weather data. These features include snowfall, temperature, rainfall, and visibility. While the quantity of these features (e.g., 20 cm of snow, 50 mm of rain, visibility of 1 km) can be "precisely" measured by sensors, human perception of the categorization of these features (e.g., "heavy snow", "heavy rain", "very poor visibility") can be subjective. This explains why we applied fuzzy logic in this key step of our algorithm to helps capture the "degree of truth" of different categories of the weather features. In our case study, snow categories are labelled in a self-explanatory manner with group increments ranging from 2-3 cm. We used the daily average snowfall. The rainfall category was labelled in a similar manner, by using the daily average rainfall. Visibility categories were based on standardized aviation visibility categories in the following manner—through the process of fuzzification—as outlined in Table II.

Defining the categories for delay times is important because different lengths of delays can have vastly different outcomes on general traffic, public transportation ridership, and rider's commute times. The lower the overall delay in rider's experience on their commute, the more is the rider's satisfaction of the public transport, which in turn leading to increases in ridership. Ridership trends are important for cities to invest money in their public transportation for future development.

TABLE I
Categorization of Time of Day Data (TOD)

Time of day (TOD)	Time range
Morning	06:30-10:30
Late morning	10:30-15:30
After work	15:30-18:30
Evening	16:30-22:30
Night	22:30-06:30

TABLE II
Categorization of Visibility (VisCat)

Fuzzy set	Visibility type	Distance range
minVal to 0.25	Very poor visibility	0-10 km
0.25 to 0.5	Poor visibility	10-20 km
0.5 to 0.75	Moderate visibility	20-30 km
0.75 to maxVal	Good visibility	> 30 km

TABLE III
Categorization of Delay Data (DelayCat)

Fuzzy set	Delay type	Time range
minVal to 0.2	No to light delay	0-5 mins
0.2 to 0.4	Light to medium delay	5-15 mins
0.4 to 0.6	Medium to high delay	15-30 mins
0.6 to 0.8	Heavy delay	30-50 mins
0.8 to maxVal	Extreme delay	> 50 mins

In this case study, we defined any delays with less than five minutes to be insignificant (No - light delay), and significant delays were further broken down into four categories—through the process of fuzzification—as outlined in Table III.

B. *Key Step 2—Frequent Pattern and Association Rule Mining*

After the data was categorized, we mine frequent patterns and association rules by using classical algorithms like Apriori [1] or FP-growth [16]. Specifically, we use FP-growth to discover frequent patterns in the form of collections of frequently co-occurring features. The following collections of fuzzified features are some samples that were observed to occur frequently in the TTC streetcar delay data:

- collections of a single fuzzified feature {"Medium-heavy wind"}.
- collection of multiple fuzzified features {"Very light snow" (2-5 cm of snow), "Poor visibility"}.

In addition, "Mechanical problem" is another frequent pattern.

Moreover, we also form interesting association rules (for associative classification) by using the aforementioned discovered frequent patterns as antecedents of the rules and the fuzzified delay type as consequents of the rules. Examples of these association rules include:

- {"Medium-heavy wind"} ⇒ "No to light delay", which reveals that medium-heavy wind is likely to lead to no (to light) delays in streetcars.
- {"Very light snow" (2-5 cm of snow), "Poor visibility"} ⇒ "Heavy delay", which reveals that very light snow when combined with poor visibility is likely to lead to heavy delays in streetcars.

C. *Key Step 3—Transit Delay Prediction with Random Forest Regression*

After the frequent patterns were discovered from the preprocessed and categorized data, we predict transit delays by using supervised learning. Specifically, taken in consideration the trade-off between prediction accuracy and explainability, we uses random forest regression to predict fuzzified delay types for streetcars:

- No (to light) delay, which ranges from 0 to 5 minutes;
- Light (to medium) delay, which ranges from 5 to 15 minutes;
- Medium (to high) delay, which ranges from 15 to 30 minutes;
- Heavy delay, which ranges from 30 to 50 minutes; and
- Extreme delay, which delays over 50 minutes.

The overall size of the input data set from the last five years was approximately 77,000 samples. A 5-fold cross validation with a 75-25% training/test split was used (meaning 75% of our data was used for training and the remaining 25% was used for testing in each fold). Using k-fold cross validation helps generalize the model and reduce over-fitting, and thus usually enhances the prediction accuracy for unseen test data. We used an ensemble size of 40 estimators when building the random forest regression model for making our prediction.

V. Evaluation

In this section, we show our evaluation on our presented fuzzy logic-based machine learning algorithm for predictive analytics on big transportation data. Specifically, we focused on three key aspects:

- data visualization
- frequent pattern mining
- prediction with random forest

A. *Visual Analytics for Data Categorization*

To get meaningful and explainable predictions, we first get an insight about the streetcar delay data by visual analytics, which provides us some valuable information about data distribution in visual means. For instance, we plotted the proportion of significant delays (i.e., delays over 5 minutes) over the total delays, for each category of snowfall, in Fig. 1, from which we observed the following:

- With 0-2 cm of snow on the ground, 59% of the total delays were significant (i.e., over 5 minutes) and the remaining 41% of time were insignificant (i.e., no delays or delays within an acceptable range of 5 minutes).
- With more snow on ground, the portion of significant delays increased (e.g., with 11-17 cm of snow on ground, the portion of significant delays went up to over 90%).
- Then, with more than 17 cm of snow on the ground, the portion of significant delays dropped from its peak (from over 90% to below 50%).

Based on the observation on Fig. 1, readers could infer the following:

- Streetcars run on the tracks (aka grooved rails) on street surfaces (i.e., pavement) and through regular traffic lights in Toronto. As snow would have an impact on general traffic, streetcars would be affected by the overall increase in traffic.
- Snow is often associated with colder temperatures, and combining snow with colder temperatures can have an impact on the machinery that streetcars rely on.
- Clearing snow can drain city resources, which can have an impact on repair crews for public transportation as well as resource allocation.
- With small amounts of snow (e.g., 0-2 cm) on the ground, delays are likely to be caused by traffic on the street shared by streetcars, other vehicles and pedestrian crossing the streets.
- With more snow on the ground (e.g., 11-17 cm), tracks may be blocked by snow and streetcars may be slowed down by other slow-moving vehicles (which slowed down due snow).
- With more than 17 cm of snow on the ground, city may have been more prepare to clean the snow (when compared with first snowfall within 17 cm of snow) and less traffic on the street (as more people may work from home). This explains why the portion of significant delays dropped from its peak of over 90% to below 50%.

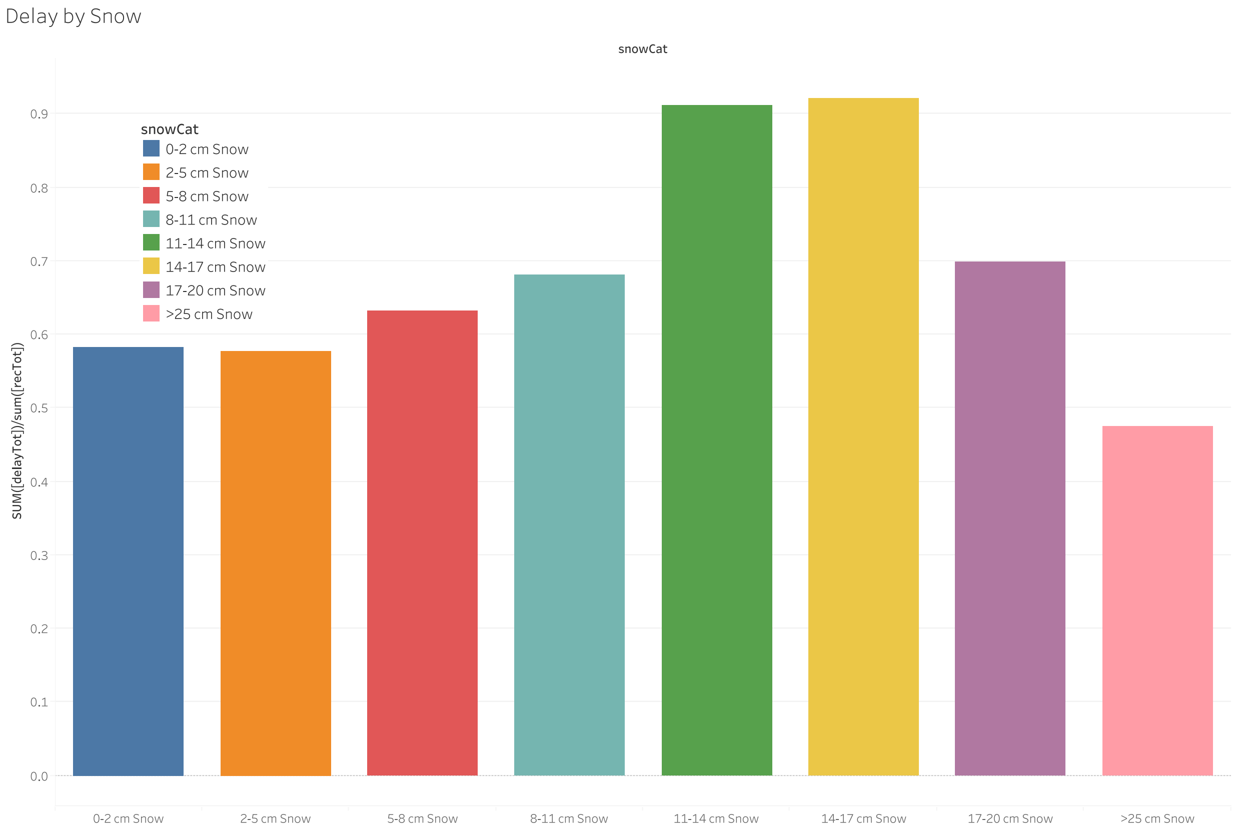

Fig. 1. Percentage of significant delays by the amounts of snow.

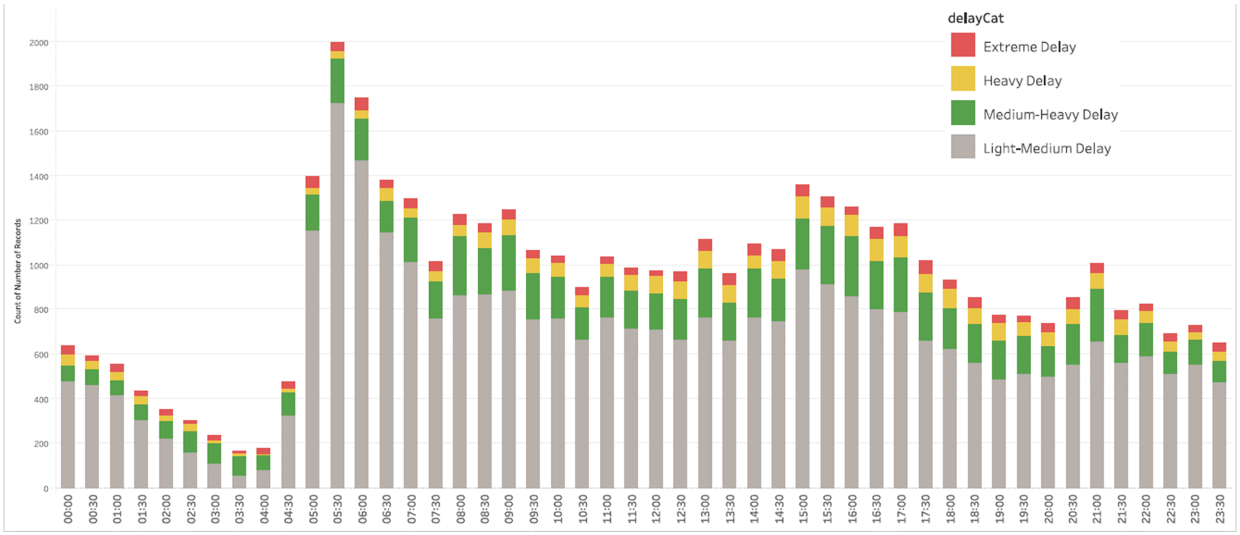

Fig. 2. Frequency of significant delays by time of day (TOD).

This evaluation result confirms our hypothesis that *snow has an impact of transportation delays*.

Next, we plotted the frequency of significant delays as a function of the time of day (TOD) in Fig. 2. It shows higher delay duration occurred around 5:30-9:00 and 15:30-18:30 in a major metropolitan city like Toronto, where people often commute early and far in the city centers. We observed:

- Increased traffic on the major streets affect the routes that streetcars take as they travel these same routes, and
- Increased ridership during peak times can have an affect on the streetcar mechanical systems.

The effect of TOD on the frequency of significant streetcar delays can be observed from the figure. Moreover, Fig. 2 helps us to divide the TOD into the five categories listed in Table I.

Furthermore, it is clear that the public transportation system is burdened before and after work during peak commute times.

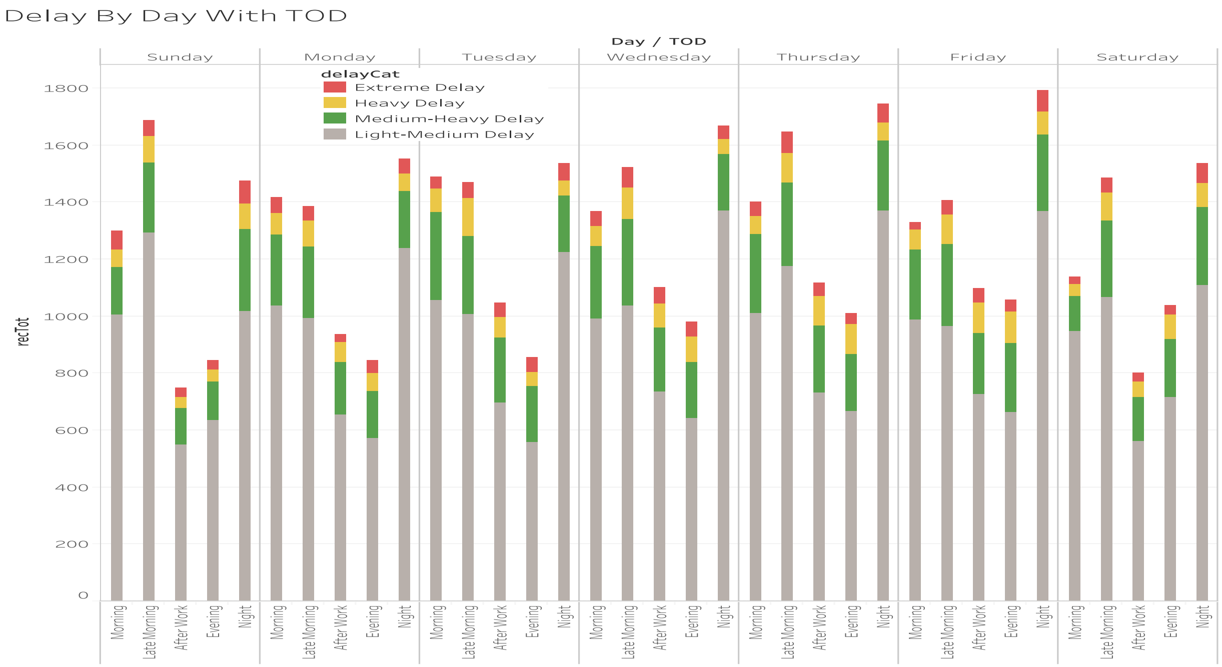

Fig. 3. Frequency of significant delays by day of week (DOW) and time categories.

The frequency of severe delays peaks between 05:30-06:30 and 15:00-17:00. The proportion of each delay category is relatively constant upon visual inspection with a slightly higher proportion for "Light (to medium) delay" during the afternoon commute when compared to the morning commute. This becomes much clear when plotting the delay by categorized TOD, grouped by day of the week (DOW), as shown in Fig. 3. The figure shows that the night (22:30-06:30) and late morning (10:30-15:30) were times when most of the significant delays (especially, light to medium delays of duration 5-15 minutes) occurred. Those are also non-peak (aka non-rush) hours when the frequency of streetcars was lower (than those in the peak hours).

In addition to examining the impact of snowfall on streetcar delays (via visual analytics), we also examined the impact of other features (e.g., temperature) on streetcar delays. The results show that, *different temperature does not have a significant impact on the significant delays*, except for the extreme cold temperature of below -25°C. Such an extreme cold more likely to attribute to mechanical breakdowns, and thus streetcar delays.

B. Mining of Frequent Delay Patterns

After the data was pre-processed and categorized, our algorithm found frequent patterns in the form of combinations of factors contributing to transit delays. Top-k frequent patterns include:

- {"Poor visibility"}, which reveals that poor visibility frequently occurs with streetcar delays (regardless of the level/severity of delay).
- {"Medium-heavy wind"}, which reveals that medium-heavy wind also frequently occurs with streetcar delays.
- {"Poor visibility", "Heavy delay"}, which reveals that poor visibility frequently occurs with heavy delays.
- {"Very light snow" (2-5 cm of snow), "Heavy delay"}, which reveals that very light snow frequently occurs with heavy delays.
- {"West direction", "Medium (to heavy) delay"}, which reveals that westbound streetcars frequently occur with medium delays.
- {"Mechanical problem", "Light (to medium) delay"}, which reveals that mechanical problems frequently occur with light delays.
- {"-5°C to 0°C", "Heavy delay"}, which reveals that temperature between -5°C to 0°C frequently occurs with heavy delays.

Then, we formed interesting association rules by using the aforementioned discovered frequent patterns as antecedents of the rules and the fuzzified delay type as consequents of the rules. Top-k interesting association rules include:

- "Poor visibility" $\Rightarrow$ "Heavy delay", which reveals that poor visibility is likely to be associated with heavy delays.
- "Very light snow" (2-5 cm of snow) $\Rightarrow$ "Heavy delay", which reveals that poor visibility is likely to be associated with heavy delays.
- "West direction" $\Rightarrow$ "Medium (to heavy) delay", which reveals that westbound streetcars are likely to be associated with medium delays.
- "Mechanical problem" $\Rightarrow$ "Light (to medium) delay", which reveals that mechanical problems are likely to be associated with light delays.
- "-5°C to 0°C" $\Rightarrow$ "Heavy delay", which reveals that temperature between -5°C to 0°C is likely to be associated with heavy delays.

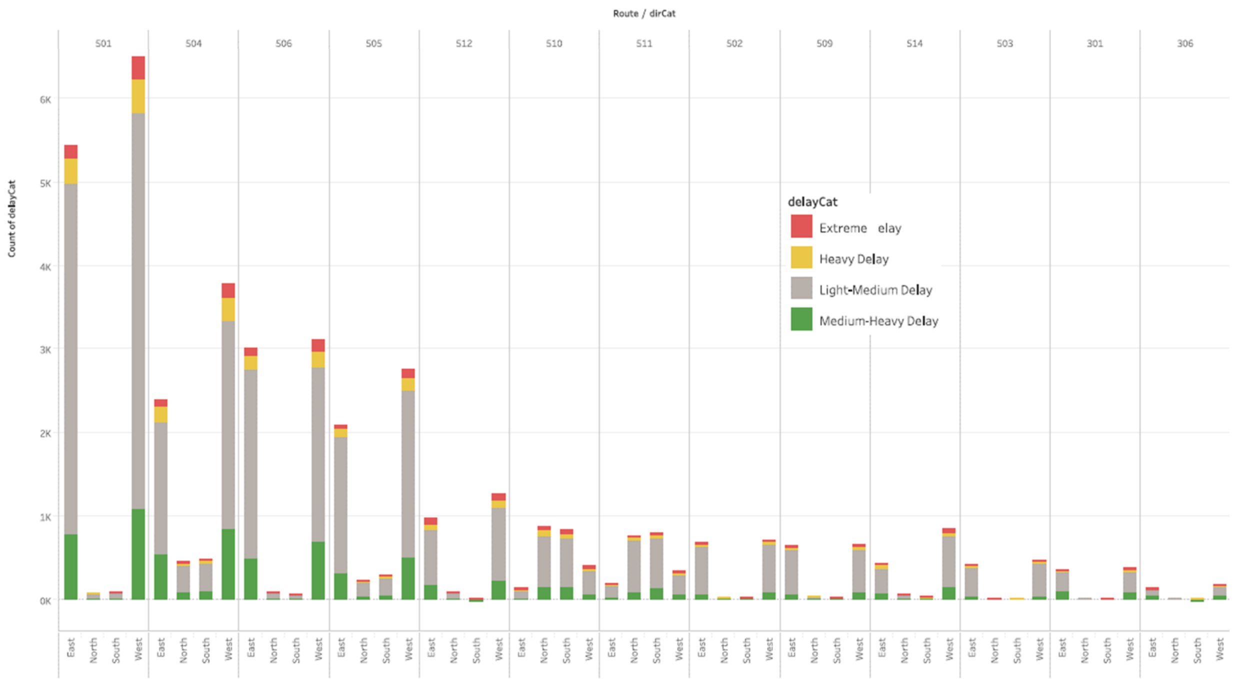

Fig. 4. Frequency of significant delays by routes and directions.

It is interesting to observe that *westbound was the only frequent pattern for the direction category*. Such an observation reveals that, a major subway line in Toronto—namely, Subway Line 1 (Yonge-University line)—runs on University Ave, which is one of the busiest North-South streets in Toronto and is slightly on the east side of the central business district of Toronto. Much more east of a bulk of the residential areas of the city. Hence, a large proportion of people transferring from the Subway Line 1 would more likely to travel westbound rather than eastbound. This means a higher burden would be placed upon streetcar routes in the westbound direction. This observation is consistent with the results shown in Fig. 4. It shows the frequency of significant delays by route directions, grouped by routes, for the top-13 frequently delayed routes among 12 regular streetcar routes (numbered 501-506, 508-512 & 514 covering 10 active, 1 suspended and 1 defunct routes) together with four all-night streetcar routes (numbered 301, 304, 306 & 310).

C. Delay Prediction with Random Forest Regression

To evaluate our algorithm, we compared it with a mean rule algorithm, which estimates the delay time based on the mean delay of the training data. Such a mean rule algorithm led to a mean prediction error (MPE) of 10.62 minutes. Note that, in metropolitan city like Toronto, commuters often require multiple transfers to get to work. An accumulative effect of 10 minute delay for each of the multiple transfers could lead to a significant loss of productive working hours (e.g., over 30 minutes for a 3-transfer trip to work, in turn causing over 5 hours per week to/from work). Moreover, many commuters often travel out of town and use public transport to get to the Union Pearson (UP) Express (which is an airport rail link running between downtown Union Station and Toronto Pearson International Airport) for fast airport travel. A delay of over 10 minutes could mean missing an UP train, causing an even further delay of around 15 minutes to wait for the next UP train, and in turn causing a missed flight.

TABLE IV
5-Fold Cross Validation with 75%-25% Training/Test Split

k-fold	Mean prediction error (MPE) in minutes
Fold 1	3.78
Fold 2	3.92
Fold 3	3.67
Fold 4	2.92
Fold 5	4.22
Mean	3.70

In contrast, when using 5-fold cross validation (in which each fold used a 75%-25% split for training/test data), the ensemble of 40 decision trees in our random forest regression led a MPE of only 3.70 minutes averaged over five folds, as shown in Table IV. Hence, delay prediction with our random forest regression led to a 70% reduction in the prediction error when compared with the mean rule algorithm. This demonstrates the usefulness and practicality of our prediction algorithm.

VI. Conclusion

Hypothesizing the weather (especially, snow) pays a role in causing delays in transit and affecting punctuality of the transit system, we designed and developed—in this paper—an innovative fuzzy logic-based machine learning algorithm for supporting predictive analytics on big transportation data. Our algorithm augments transit data (in particular, streetcar delay data) with weather information, pre-processes them with fuzzy-logic based categorization, visualizes and analyzes the augmented data and their characteristics, mines frequent patterns (i.e., collections of frequently co-occurring features that contributing to the delays) and interesting association rules (which can be applicable for associative classification), and predicts delays with a random forest. The algorithm takes into consideration the trade-off between prediction accuracy and explainability, and leads to a reasonable mean prediction error (MPE) of 3.70 minutes averaged over a 5-fold cross validation. Such a result is promising toward development of a predictive intelligent transport system (ITS).

In the current paper, we illustrated our algorithm with a case study on real-life transportation data focusing the streetcar delays in the Canadian city of Toronto. As ongoing and future work, we are transferring our knowledge (via transfer learning) and adapting our algorithm to other modes of transit in Toronto (e.g., subway, bus), as well as other cities. Moreover, we are also exploring fuzzification on additional features, as well as incorporating alternative data mining and machine learning techniques and/or other approaches (e.g., OLAP [7], [10]–[13]), to further enhance predictive analytics on big transportation data. To a further extent, we plan to apply transfer learning [22] for supporting predictive analytics on big epidemic data like COVID-19 data.

Acknowledgment

This project is partially supported by NSERC (Canada) and University of Manitoba.

References

[1] R. Agrawal & R. Srikant, "Fast algorithms for mining association rules," in VLDB 1994, pp. 487-499.

[2] S. Ahn, S. V. Couture, A. Cuzzocrea, K. Dam, G. M. Grasso, C. K. Leung, K. L. McCormick, & B. H. Wodi, "A fuzzy logic based machine learning tool for supporting big data business analytics in complex artificial intelligence environments, in FUZZ-IEEE 2019, pp. 1259-1264.

[3] A. A. Audu, A. Cuzzocrea, C. K. Leung, K. A. MacLeod, N. I. Ohin, & N. C. Pulgar-Vidal, "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] L. Bellatreche, C. K. Leung, Y. Xia, & D. Elbaz, "Advances in cloud and big data computing," CCPE 31(2): e5053:1-e5053:3 (2019)

[5] L. Biacino and G. Gerla, "Fuzzy logic, continuity and effectiveness," Archive for Mathematical Logic 41(7): 643-667 (2002)

[6] C. Blundo, C. de Maio, M. Parente, & L. Siniscalchi, "An intelligent and private method to profile social network users," in FUZZ-IEEE 2019, pp. 900-905.

[7] G. Chatzimilioudis, A. Cuzzocrea, D. Gunopulos, & N. Mamoulis, "A novel distributed framework for optimizing query routing trees in wireless sensor networks via optimal operator placement," JCSS 79(3): 349-368 (2013)

[8] S. I. Chien & C. M. Kuchipudi, "Dynamic travel time prediction with real-time and historic data," ASCE JTE 129(6): 608-616 (2003)

[9] J. de Guia, M. Devaraj, & C. K. Leung, "DeepGx: deep learning using gene expression for cancer classification," in IEEE/ACM ASONAM 2019, pp. 913-920.

[10] A. Cuzzocrea, "Combining multidimensional user models and knowledge representation and management techniques for making web services knowledge-aware," WIAS 4(3): 289-312 (2006)

[11] A. Cuzzocrea & E. Bertino, "Privacy preserving OLAP over distributed XML data: a theoretically-sound secure-multiparty-computation approach," JCSS 77(6): 965-987 (2011)

[12] A. Cuzzocrea, R. Moussa, & G. Xu, "OLAP*: effectively and efficiently supporting parallel OLAP over big data," in MEDI 2013, pp. 38-49.

[13] A. Cuzzocrea & V. Russo, "Privacy preserving OLAP and OLAP security," Encyclopedia of Data Warehousing and Mining, 2e, pp. 1575-1581. (2009)

[14] D. Deng, J. J. Mai, C. K. Leung, & A. Cuzzocrea, "Cognitive-based hybrid collaborative filtering with rating scaling on entropy to defend shilling influence," in ICNCC 2019, pp. 176-185.

[15] J. Y. Halpern, Reasoning about Uncertainty. MIT Press (2003)

[16] J. Han, J. Pei, & Y. Yin, "Mining frequent patterns without candidate generation," in ACM SIGMOD 2000, pp. 1-12.

[17] M. Khan, N. Javaid, M. N. Iqbal, M. Bilal, S. F. A. Zaidi, & R. A. Raza, "Load prediction based on multivariate time series forecasting for energy consumption and behavioral analytics," in CISIS 2018, pp. 305-316.

[18] M. Kormáksson, L. Barbosa, M. R. Vieira, & B. Zadrozny, "Bus travel time predictions using additive models," in IEEE ICDM 2014, pp. 875-880.

[19] E. Kulla, S. Morita, K. Katayama, & L. Barolli, "Route lifetime prediction method in VANET by using AODV routing protocol (AODV-LP)," in CISIS 2018, pp. 3-11.

[20] L. V. S. Lakshmanan, C. K. Leung, & R. T. Ng, "The segment support map: Scalable mining of frequent itemsets," ACM SIGKDD Explorations 2(2): 21-27 (2000)

[21] C. K. Leung, P. Braun, C. S. H. Hoi, J. Souza, & A. Cuzzocrea, "Urban analytics of big transportation data for supporting smart cities," in DaWaK 2019, pp. 24-33.

[22] C. K. Leung, A. Cuzzocrea, J. J. Mai, D. Deng, & F. Jiang, "Personalized DeepInf: enhanced social influence prediction with deep learning and transfer learning," IEEE BigData 2019, pp. 2871-2880.

[23] C. K. Leung, S. K. Tanbeer, & J. J. Cameron, "Interactive discovery of influential friends from social networks," SNAM 4(1): 154:1-154:13 (2014)

[24] C. K. Leung & Y. Zhang, "An HSV-based visual analytic system for data science on music and beyond," IJACDT 8(1): 68-83 (2019)

[25] Y. Lin, X. Yang, N. Zou, & L. Jia, "Real-time bus arrival time prediction: case study for Jinan," ASCE TE 139(11), 1133-1140 (2013)

[26] J. Liu, Z. Chang, C. K. Leung, R. C. W. Wong, Y. Xu, & R. Zhao, "Efficient mining of extraordinary patterns by pruning and predicting," ESWA 125: 55-68 (2019)

[27] K. J. Morris, S. D. Egan, J. L. Linsangan, C. K. Leung, A. Cuzzocrea, & C. S. H. Hoi, "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.

[28] P. Rajput, D. Toshniwal, & A. Agggarwal, "Improving infrastructure for transportation systems using clustering," in BDA 2017, pp. 129-143.

[29] O. A. Sarumi & C K. Leung, "Exploiting anti-monotonic constraints for mining palindromic motifs from big genomic data," in IEEE BigData 2019, pp. 4864-4873.

[30] A. Shalaby & A. Farhan, "Prediction model of bus arrival and departure times using AVL and APC data," J. Public Transp. 7(1): 41-61 (2004)

[31] S. P. Singh, C. K. Leung, F. Jiang, & A. Cuzzocrea, "A theoretical approach to discover mutual friendships from social graph networks," in iiWAS 2019, pp. 212-221.

[32] J. Souza, C. K. Leung, & A. Cuzzocrea, "An innovative big data predictive analytics framework over hybrid big data sources with an application for disease analytics," AINA 2020, pp. 669-680.

[33] D. Sun, H. Luo, F. Fu, W. Liu, X. Liao, & M. Zhao, "Predicting bus arrival time on the basis of global positioning system data," SAGE TRR 2034(1): 62-72 (2007)

[34] S. K. Tanbeer, C. K. Leung, & J. J. Cameron, "Interactive mining of strong friends from social networks and its applications in e-commerce," JOCEC 24(2-3): 157-173 (2014)

[35] L. Vanajakshi, S. C. Subramanian, & R. Sivanandan, "Travel time prediction under heterogeneous traffic conditions using global positioning system data from buses," IET-ITS 3(1): 1-9 (2009)

[36] B.M. Williams & L. A. Hoel, "Modeling and forecasting vehicle traffic flow as a seasonal ARIMA process: theoretical basis and empirical results," ASCE JTE 129(6): 664-672 (2003)

[37] L. A. Zadeh, "Fuzzy sets," Information and Control 8(3): 338-353 (1965)