Vol. 15 No. 3 (2024):239-246
DOI: https://doi.org/10.61841/turcomat.v15i3.14797

DATAFITS: A HETEROGENEOUS DATA FUSION FRAMEWORK FOR TRAFFIC AND INCIDENT PREDICTION

[1]Dr. Y. Geetha Reddy, [2]P. Sindhoora,[3]R. Sharanya,[4]Y. Swathi
[1]Professor,[234]Students
Department Of CSE
Malla Reddy Engineering College for Women

ABSTRACT

In order to create a complete dataset, this study presents DataFITS (Data Fusion on Intelligent Transportation System), an open-source system that gathers and fuses traffic-related data from several sources. Our hypothesis is that traffic models may benefit from improved information coverage and quality thanks to a heterogeneous data fusion architecture, which would boost the effectiveness and dependability of ITS systems. Two applications that made use of event categorization and traffic estimate models confirmed our hypothesis. For nine months, DataFITS gathered four different kinds of data from seven different sources and combined them into a spatiotemporal domain. While incident categorization utilized the k-nearest neighbors (k-NN) method with Dynamic Time Warping (DTW) and Wasserstein metric as distance measurements, traffic estimation models used polynomial regression and descriptive statistics. The findings show that by fusing data, DataFITS was able to enhance information quality for up to 40% of all roads and dramatically expand road coverage by 137%. While incident classification reached 90% accuracy on binary tasks (incident or non-incident) and about 80% on categorizing three distinct categories of events (accident, congestion, and non-incident), traffic estimate earned an R2 score of 0.91 using a polynomial regression model.

1. INTRODUCTION

In order to improve mobility and safety for both people and products, current Intelligent Transportation Systems (ITSs) rely heavily on data availability in their design. ITSs use models to better comprehend the varied patterns of the transportation system [1]. The relevance of these systems has grown significantly in recent years due to the heavy reliance of contemporary civilization on dependable and efficient transportation. Both the number of automobiles registered and the number of passengers carried by public transportation have significantly increased in Germany alone, hitting all-time highs of 48.5 million cars in 2022 and 12.7 billion people transported in 2019 (before to the pandemic) [2], [3]. Urban regions therefore see a rise in time delays, pollution, fuel

Vol. 15 No. 3 (2024):239-246
DOI: https://doi.org/10.61841/turcomat.v15i3.14797

consumption, and traffic-related events (such as accidents and congestion) [4].

Because of this, attempts to develop the next generation of transportation systems that are economical, environmentally friendly, and powered by data analysis and communication technologies have been spearheaded by academics and industry. Our hypothesis is that the coverage and quality of data used as input for traffic models may be improved by a heterogeneous data fusion framework, which would increase the effectiveness and dependability of ITS systems. In light of this, we present the Data Fusion on Intelligent Transportation System (Data FITS) framework, which offers a spatiotemporal data fusion for training models in two ITS applications: incident categorization and traffic estimate. Real heterogeneous data (such as weather, traffic, and incidents) are gathered and combined by Data FITS from a variety of sources (such as open databases and map apps). The data is then prepared by correcting mistakes and modifying the data structure before being fused at the precise place and time. By proposing two ITS applications and using data characterisation to measure the advantages of merging diverse data sources, our prediction is confirmed. The efficacy of both apps in traffic estimation and incident classification validates the advantages of increased data coverage and quality. Therefore, the primary contributions of this study are:

• Data FITS is an open-source system for heterogeneous spatiotemporal data fusion that is accessible in a public code repository and covers data gathering, processing, and fusion.1.

• The description of a heterogeneous dataset that includes actual traffic data from two German cities that was gathered over a nine-month period from seven different sources and made available with the repository.

• A comparison between single and fused datasets; two traffic estimate models, one based on descriptive statistics and the other on polynomial regression with various characteristics including time, kind of road, and weather.

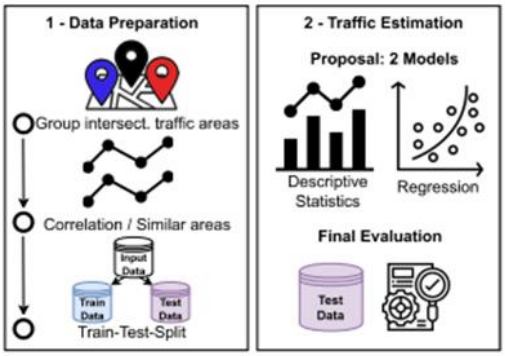
• An incident classification model using Wasserstein and Dynamic Time Warping (DTW) as distance techniques, trained and assessed on heterogeneous fused data using k-nearest neighbors (k-NN).

This is how the remainder of the paper is structured. Section II examines and contrasts our method with contemporary literature that uses data fusion to create applications such as traffic prediction and event categorization. Section III describes the traffic data applications and the Data FITS architecture. In order to validate our hypothesis, Section IV assesses the efficiency of our traffic estimating and incident categorization models as well as the

Vol. 15 No. 3 (2024):239-246
DOI: https://doi.org/10.61841/turcomat.v15i3.14797

performance of our framework utilizing the heterogeneous fused data. In Section V, we finally wrap up this work by emphasizing unresolved issues that need more research.

2. SYSTEM ARCHITECTURE



3. EXISTING SYSTEM

Significant data from actual or virtual sensors is needed to construct ITS applications [5]. A platform for gathering, processing, and exporting heterogeneous data from smart city sensors is presented by Vitor et al. [4], who also provide various statistics and visualizations. But the focus of their platform is data security. In a similar vein, [6] suggests creating a smart city data platform using data from several cities. Unlike our approach, we concentrate on fusing data to improve both the amount and quality of the information, and we evaluate the benefits of employing fused data via two applications for ITS. By merging data from many sources, data fusion improves spatiotemporal information [7], [8], [9], and [10]. Data fusion is useful for several applications, including route planning [12] and emergency management [11]. However, extra preprocessing is needed to fuse heterogeneous data in order to merge different characteristics and data formats [13], [14]. The two applications—traffic estimates and incident classification—that are enabled by data fusion are the subject of this inquiry, together with the strategies used to meet their objectives, including data collecting, fusion, machine learning, correlation, and the use of various data types.

One essential smart city application for improved transportation management is traffic estimation. The goal of this study is to provide accurate and trustworthy traffic estimate utilizing historical data by examining data fusion, spatiotemporal correlation, and machine learning algorithms. Big data presents an opportunity for heterogeneous data fusion because of the growing amount of traffic-related data that is collected by commercial apps (such as Bing and Google Maps) and open databases maintained by governmental bodies. [15]. Combining stationary sensor data—like loop detectors or traffic cameras—with information from probing vehicles—like cameras, GPS, cellphone data, or vehicular sensors—presents a difficulty. Anand et al. [16] improved a traffic estimate method by fusing journey time (from GPS) and traffic

Vol. 15 No. 3 (2024):239-246
DOI: https://doi.org/10.61841/turcomat.v15i3.14797

flow data (from cameras) using a Kalman filter. Machine Learning (ML) is used in several of the more current traffic prediction models [17], [18], [19], [20], [21], [22], [23], [24], [25]. An auto-regressive model that utilizes data from a traffic simulator and adjusts to occurrences such as accidents is suggested in reference [17].

According to their findings, there is a 12% mistake in estimations made up to 30 minutes in advance. In the meanwhile, [18] uses deep learning algorithms to estimate traffic, demonstrating an increase in efficiency and accuracy. These techniques talk about using machine learning (ML) to build precise traffic estimate models, however they don't take into account other approaches like data fusion, correlation, etc.

Spatial-temporal correlation is a tool used by several machine learning techniques to enhance traffic estimate quality. A neural network (NN)-based estimate using the Gated Recurrent Unit (GRU) and Graph Convolutional Network (GCN) models is suggested in [19] and is available to the general public. Road network geographical dependencies are captured by the GCN, while temporal dependencies are captured by the GRU, which also detects dynamic changes in traffic data. Analogously, other NN-based methods [20] and [21] demonstrate comparable gains in accuracy via data correlation. An open-source deep learning framework using GCN is proposed by Wang et al. [22] to predict network-wide traffic many steps in advance. Another open-source approach, the Graph Multi Attention Network (GMAN), is introduced by Zheng et al. [23]. It provides long-term traffic estimate up to one hour in advance utilizing an encoder-decoder design. These methods do not provide a way to gather or combine data; instead, they use correlation to enhance the models that have been presented and provide access to their data. Similar to our approach, a limited body of work uses ML, spatiotemporal correlation, and data fusion to estimate traffic. The authors of [26] combine traffic data from both dynamic and stationary sensors, taking into account the spatiotemporal link between road segment traffic levels.

The fused data is processed using a Multiple Linear Regression (MLR) model to improve the accuracy of traffic estimates. This method, in contrast to ours, ignores various data kinds and sources and just uses traffic data from sensors. In order to improve traffic estimates, Zhao et al. [24] provide a generic framework for spatiotemporal data fusion. The strategy combines direct and indirect traffic-related data as input for two distinct ML models, introducing a fusion mechanism to increase accuracy. The weather and place of interest information found in the indirect traffic-related data characteristics is employed to enhance the

Vol. 15 No. 3 (2024):239-246
DOI: https://doi.org/10.61841/turcomat.v15i3.14797

estimate quality. While the authors in [24] take into account locations of interest and meteorological conditions, our work concentrates on incident-related data, and their model employs pre-existing datasets, providing no means for data gathering.

Disadvantages

- The data fusion framework Data FITS and data applications traffic estimate and event categorization were not implemented by the system.
- The combined data from DataFITS is categorized into traffic zones with one or more road segments without being cleansed or stripped of incident-related information since the model does not call for it.

4. PROPOSED SYSTEM

The system provides a spatiotemporal fusion of data used to train models for two ITS applications, traffic estimate and incident categorization. It does this by proposing the Data Fusion on Intelligent Transportation System (DataFITS) architecture. Real heterogeneous data (such as weather, traffic, and incidents) are gathered and combined by DataFITS from a variety of sources (such as open databases and map apps). DataFITS then fixes mistakes, modifies the data structure, and ultimately fuses the data in the precise place and time. By proposing two ITS applications and using data characterisation to measure the advantages of merging diverse data sources, our prediction is confirmed. The two apps' performance validates the advantages of greater data coverage and quality when it comes to traffic estimation and incident classification.

Advantages

• DataFITS, an open-source framework for heterogeneous spatiotemporal data fusion, is accessible in a public code repository and covers data gathering, processing, and fusion.

• The description of a heterogeneous dataset that includes actual traffic data from two German cities that was gathered over a nine-month period from seven different sources and supplied with the repository.

• A comparison of single and fused datasets, as well as two traffic prediction models—one based on descriptive statistics and the other on polynomial regression with various factors including time, kind of road, and weather.

• Dynamic Time Warping (DTW) and Wasserstein distance approaches were used in conjunction with k-nearest neighbors (k-NN) to train and assess an event classification model using heterogeneous fused data.

5. IMPLEMENTATION

Modules description

Vol. 15 No. 3 (2024):239-246
DOI: https://doi.org/10.61841/turcomat.v15i3.14797

Service Provider

The Service Provider must provide a valid user name and password to log in to this module. He can do some tasks after logging in successfully, such browsing datasets and training and testing data sets. View Traffic Incident Type, View Traffic Incident Type Ratio, Download Predicted Data Sets, View Traffic Incident Type Ratio Results, View All Remote Users, and View Trained and Tested Accuracy in Bar Chart.

View and Authorize Users

The administrator may see a list of all enrolled users in this module. The administrator may see user information here, including name, email address, and address, and they can also approve people.

Remote User

There are n numbers of users present in this module. Prior to beginning any actions, the user must register. The user's information is saved in the database when they register. Upon successful registration, he must use his permitted user name and password to log in. Following a successful login, the user may do certain tasks including VIEW YOUR PROFILE, PREDICT TRAFFIC INCIDENT TYPE, and REGISTER AND LOGIN.

6. CONCLUSION

In this work, we provide Data FITS, an open-source framework for data fusion that gathers, examines, and combines disparate data. Our hypothesis is that heterogeneous data fusion improves datasets for ITS applications by increasing both the number and quality of data. We created two ITS programs to confirm this: one classified events as accidents, congestion, or non-incidents using traffic and incident data, while the other utilized polynomial regression to predict traffic levels. By creating a fused dataset, we were able to quantify the benefits of Data FITS using actual heterogeneous data from two German cities. According to our findings, Data FITS combined information from various sources for 40% of all roads, resulting in a 137% increase in total road coverage. Furthermore, the polynomial regression-based traffic estimating model performed better than our descriptive statistics-based prior method, obtaining a high R2 score of 0.91, low error metrics of 0.05, and accurate traffic predictions utilizing the fused dataset. While the spatiotemporal coverage of the predicted regions was significantly increased, the fused dataset estimation only slightly improved accuracy when compared to utilizing a single sources dataset. Our event classification model, which we evaluated, achieves a 90% binary classification accuracy rate by combining traffic and

Vol. 15 No. 3 (2024):239-246
DOI: https://doi.org/10.61841/turcomat.v15i3.14797

incident data. Preprocessing the data, which included eliminating hazy traffic patterns, increased accuracy by 29% on average. A little lower accuracy of 86% was obtained from the categorization of occurrences into distinct categories; F1 scores showed differential performance across classes. In order to address this issue, we oversampled the training dataset, which produced a more consistent data representation and an accuracy rate of 80% for each class. Additional accident data collection may also help to resolve this issue. Through the collection and fusion of more data kinds, enhancement of data quality and performance, and expansion of data analysis, we want to grow the Data FITS framework. We concentrate on data types including photos and social media, which call for techniques like image processing and natural language processing (NLP). Our goal is to compare existing models with alternative models and hyper-parameters for ITS applications using automated machine learning. Additionally, we want to examine and include the association between events and traffic in the models used to estimate traffic. Furthermore, our aim is to investigate the use of big data in military contexts, merging data from both military and civilian domains to facilitate strategic maneuvers in urban combat. In order to do this, our system may be improved to gather and integrate various forms of information (text, image) to provide common operational images and validate/verify information, preventing false information from influencing political choices.

REFERENCES

[1] L. Zhu, F. R. Yu, Y. Wang, B. Ning, and T. Tang, “Big data analytics in intelligent transportation systems: A survey,” IEEE Trans. Intell. Transp. Syst., vol. 20, no. 1, pp. 383–398, Jan. 2019.

[2] Umweltbundesamt. (2022). Verkehrsinfrastruktur und fahrzeugbestand. Accessed: Dec. 12, 2022. [Online]. Available:https://www.umweltbundesamt.de/daten/verkehr/verkehrsinfrastrukturfahrzeugbestand

[3] German Federal Statistical Office (Destatis). (2022). Passengers Carried in Germany. Accessed: Jul. 12, 2022. [Online]. Available: https://www.destatis.de/EN/Themes/Economic-Sectors-Enterprises/Transport/Passenger-Transport/Tables/passengerscarried.html

[4] G. Vitor, P. Rito, and S. Sargento, “Smart city data platform for real-time processing and data sharing,” in Proc. IEEE Symp. Comput. Commun.(ISCC), Sep. 2021, pp. 1–7.

[5] A. B. Campolina, P. H. L. Rettore, M. Do Val Machado, and A. A. F. Loureiro, “On the design of vehicular virtual sensors,” in Proc. 13th Int. Conf. Distrib. Comput. Sensor Syst. (DCOSS), Jun. 2017, pp. 134–141.

Vol. 15 No. 3 (2024):239-246
DOI: https://doi.org/10.61841/turcomat.v15i3.14797

[6] S. Jeong, S. Kim, and J. Kim, "City data hub: Implementation of standard-based smart city data platform for interoperability," Sensors, vol. 20, no. 23, p. 7000, Dec. 2020. [Online]. Available: https://www.mdpi.com/1424-8220/20/23/7000

[7] L. Zhang, Y. Xie, L. Xidao, and X. Zhang, "Multi-source heterogeneous data fusion," in Proc. Int. Conf. Artif. Intell. Big Data (ICAIBD), May 2018, pp. 47–51.

[8] P. H. L. Rettore, B. P. Santos, A. B. Campolina, L. A. Villas, and A. A. F. Loureiro, "Towards intra-vehicular sensor data fusion," in Proc. IEEE 19th Int. Conf. Intell. Transp. Syst. (ITSC), Nov. 2016, pp. 126–131.

[9] P. H. L. Rettore, A. B. Campolina, L. A. Villas, and A. A. F. Loureiro, "A method of eco-driving based on intra-vehicular sensor data," in Proc. IEEE Symp. Comput. Commun. (ISCC), Jul. 2017, pp. 1122–1127.

[10] P. H. L. Rettore, A. B. Campolina, A. Souza, G. Maia, L. A. Villas, and A. A. F. Loureiro, "Driver authentication in VANETs based on intravehicular sensor data," in Proc. IEEE Symp. Comput. Commun. (ISCC), Jun. 2018, pp. 00078–00083.

[11] G. L. Foresti, M. Farinosi, and M. Vernier, "Situational awareness in smart environments: Socio-mobile and sensor data fusion for emergency response to disasters," J. Ambient Intell. Humanized Comput., vol. 6, no. 2, pp. 239–257, Apr. 2015.

[12] H. Wen, Y. Lin, and J. Wu, "Co-evolutionary optimization algorithm based on the future traffic environment for emergency rescue path planning," IEEE Access, vol. 8, pp. 148125–148135, 2020.

[13] P. H. Rettore, G. Maia, L. A. Villas, and A. A. F. Loureiro, "Vehicular data space: The data point of view," IEEE Commun. Surveys Tuts., vol. 21, no. 3, pp. 2392–2418, 3rd Quart., 2019.

[14] S. A. Kashinath et al., "Review of data fusion methods for realtime and multi-sensor traffic flow analysis," IEEE Access, vol. 9, pp. 51258–51276, 2021.

[15] W. Jiang and J. Luo, "Big data for traffic estimation and prediction: A survey of data and tools," Appl. Syst. Innov., vol. 5, no. 1, p. 23, Feb. 2022.