PIONEER
ACADEMIC PUBLISHING LIMITED

Journal of Progress in Engineering and Physical Science
ISSN 2709-4006
www.pioneerpublisher.com/jpeps
Volume 3 Number 3 September 2024

Data Analysis and Congestion Prediction Model in Intelligent Transportation Systems

Wei Song[1]

[1] Beijing Luoqi Yunda Technology Co., Ltd., Beijing 100041, China
Correspondence: Wei Song, Beijing Luoqi Yunda Technology Co., Ltd., Beijing 100041, China.

doi:10.56397/JPEPS.2024.09.01

Abstract

The relentless march of urbanization has led to a surge in population density and a dramatic increase in the number of motor vehicles, with traffic congestion emerging as a significant barrier to urban development. Intelligent Transportation Systems (ITS) utilize advanced information technology to monitor and analyze traffic flow in real time, playing a pivotal role in alleviating urban traffic pressures and enhancing road usage efficiency. This paper aims to propose a data analysis method based on machine learning algorithms for real-time traffic monitoring and congestion prediction, thereby providing scientific decision support for urban traffic management.

The research encompasses methods of traffic data collection, including data on vehicle location, speed, and flow; preprocessing and feature extraction techniques for traffic data to improve data quality and extract features useful for congestion prediction; and the application of various machine learning algorithms to establish a traffic congestion prediction model. Additionally, this paper assesses and optimizes the model and tests it with actual traffic data to verify its effectiveness and practicality.

The research findings indicate that the proposed machine learning-based method can effectively predict traffic congestion, providing a powerful tool for traffic management departments, aiding in preemptive measures to reduce traffic delays and improve the travel experience for citizens. This study not only enriches the field of intelligent transportation systems research but also provides theoretical foundations and technical support for urban traffic management practices.

Keywords: intelligent transportation systems, traffic data analysis, congestion prediction, machine learning, decision support

1. Introduction

Urbanization is a defining feature of modern societal development, and with it comes a concentration of population and a sharp increase in the number of motor vehicles. The acceleration of urbanization has brought about numerous challenges, with traffic congestion being particularly pronounced. It not only increases the time and cost of commuting but also exacerbates environmental pollution and energy consumption. As urban areas continue to expand and transportation demands grow, traditional traffic management systems are struggling to meet the needs of modern cities, and the issue of traffic congestion is in urgent need of resolution.

Intelligent Transportation Systems (ITS) have emerged in response, integrating advanced information technology, data communication and transmission technology, electronic sensing technology, computer processing technology, and systems engineering. They are designed to achieve real-time monitoring and management of road traffic, improve road usage efficiency, and ensure traffic safety. The development of ITS is of significant importance for alleviating traffic pressures and enhancing the level of urban traffic management.

Data analysis plays a central role in traffic management. By collecting, processing, and analyzing a vast amount of traffic data, patterns and trends in traffic flow can be revealed, providing a scientific basis for traffic planning and management. Data-driven decision support systems can assist traffic managers in optimizing traffic signal control, predicting and alleviating traffic congestion, and planning traffic infrastructure construction, thereby improving the operational efficiency of the traffic system.

The motivation for this study lies in utilizing data analysis and machine learning technology to construct an efficient traffic congestion prediction model that achieves real-time monitoring and early warning of urban traffic conditions. By accurately predicting the occurrence of traffic congestion, traffic management departments can take preemptive measures, such as adjusting traffic signals, issuing travel advice, and guiding traffic flows, to reduce traffic delays and improve the travel experience for citizens.

The purpose of the research is to explore and achieve the following objectives:

- Collect and analyze urban traffic data to identify patterns and trends in traffic flow.
- Develop a traffic congestion prediction model based on machine learning to enhance the accuracy and reliability of predictions.
- Evaluate the model's performance in actual traffic systems to provide decision support for traffic management.
- Optimize model performance and explore the applicability and flexibility of the model under different traffic environments and conditions.

Through this research, it is expected to provide new perspectives and technical support for the construction and development of intelligent transportation systems and offer scientific and effective solutions for urban traffic management.

2. Related Work

Intelligent Transportation Systems (ITS) have become a hot topic in the fields of traffic engineering and urban planning in recent years as a key technology to address the challenges of urbanization. ITS integrates various high-tech methods, such as sensors, surveillance cameras, GPS positioning systems, and wireless communication technology, achieving real-time monitoring and intelligent management of urban traffic flow. With the development of big data and artificial intelligence technologies, research and applications of ITS have expanded from single traffic monitoring to multiple aspects including traffic data analysis, prediction, and management decision-making.

In the field of traffic data analysis, researchers have developed various methods to process and analyze massive traffic data. These methods include data cleaning, feature engineering, and pattern recognition, aiming to extract valuable information from complex data to provide decision support for traffic management. For example, by analyzing vehicle speed and flow, traffic bottlenecks and accident-prone areas can be identified, thereby optimizing traffic signal control strategies.

As one of the core functions of ITS, congestion prediction has attracted a lot of research attention. There are various types of congestion prediction models, including time series analysis models, machine learning models, and deep learning models. Time series analysis methods, such as ARIMA models, rely on historical traffic data to predict future traffic flow. Machine learning models, including Support Vector Machine (SVM), Random Forest (RF), and Gradient Boosting Decision Tree (GBDT), predict by learning the relationship between traffic data features and congestion. In recent years, deep learning models, such as Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN), have received more and more attention due to their advantages in processing sequential data.

Despite the fact that existing congestion prediction models have achieved good results in

some scenarios, they still face challenges in terms of accuracy, generalization ability, and real-time performance. For example, traditional machine learning models may require a lot of feature engineering and have limited ability to capture nonlinear relationships. Although deep learning models perform well in automatic feature extraction, they often require a large amount of training data and computing resources. In addition, the traffic system has a high degree of dynamics and uncertainty. Factors such as weather changes and special events can affect traffic flow, which poses higher requirements for the robustness of congestion prediction models.

This study aims to comprehensively consider the achievements and limitations of existing research, and proposes a traffic data analysis method based on machine learning algorithms to construct an efficient and accurate traffic congestion prediction model. Through comparative analysis of existing prediction models, this study will explore model architectures suitable for different traffic environments and conditions, optimize model performance, and verify its effectiveness and practicality through integration and testing in actual traffic systems.

3. Traffic Data Collection Methods

In Intelligent Transportation Systems (ITS), accurate and reliable traffic data collection is the foundation for effective traffic monitoring and congestion prediction. Data sources are diverse, including key information such as vehicle location, speed, and flow, which form the core input for traffic analysis and prediction.

Vehicle location data is usually obtained through the Global Positioning System (GPS), which can provide the precise latitude and longitude information of vehicles. Speed data can be calculated through GPS devices or measured using loop sensors, radar, or laser speed measurement devices on the road. Flow data reflects the number of vehicles passing a certain section of the road within a specific period of time and can be statistically obtained through cameras, infrared sensors, or ultrasonic sensors.

The advancement of data collection technology has greatly enriched the types and quality of traffic data. For example, advanced cameras can not only capture vehicle images but also identify vehicle types, calculate vehicle speeds, and estimate traffic flow through image processing technology. In addition, as a mature tool for speed and flow monitoring, the magnetic induction loop is widely used in traffic data collection due to its high accuracy and stability.

However, the data collection process also faces some challenges. The first is the issue of data integrity. Due to equipment failure, bad weather, or line-of-sight obstruction, data may be missing or inaccurate. The second is the issue of data real-time performance. Traffic conditions change rapidly, and the delay in data collection and transmission may affect the timeliness of the prediction model. In addition, privacy protection is also an important consideration, especially when using devices such as cameras that may involve personal privacy.

To address these challenges, researchers and engineers have adopted a variety of solutions. Data fusion technology improves the integrity and reliability of data by integrating data from different sources. Wireless communication technologies such as Wi-Fi, 4G/5G, etc., ensure that data can be quickly transmitted to the monitoring center. In terms of privacy protection, measures such as data anonymization processing and access control can be used to protect personal privacy while effectively monitoring traffic.

In summary, traffic data collection is the foundation for building Intelligent Transportation Systems. By continuously optimizing data collection technology and improving data accuracy and real-time performance, it can effectively support real-time traffic monitoring and congestion prediction, providing strong data support for urban traffic management.

4. Traffic Data Preprocessing and Feature Extraction

In the data analysis process of Intelligent Transportation Systems, preprocessing and feature extraction of traffic data are crucial steps that directly affect the quality of subsequent analysis and the performance of predictive models.

Firstly, data cleaning is the initial step in preprocessing, aimed at removing invalid, erroneous, or incomplete data records to ensure the quality of the dataset. This includes handling missing values, outliers, and duplicate records. For instance, imputing missing speed data through interpolation or mean substitution, or

identifying and addressing abnormal speed peaks using box plots. Data standardization involves scaling data of different magnitudes and units to the same range, eliminating scale differences between variables, and improving the stability and convergence rate of model training.

Next, feature selection and construction are key to building an effective predictive model. Feature selection identifies the most useful features from raw data for traffic congestion prediction, such as vehicle speed, flow, timestamps, day of the week, weather conditions, etc. Feature construction generates new features through mathematical transformations or combinations of multiple original features, such as calculating the rate of change in speed between adjacent time points to reflect trends in traffic flow. These features not only provide richer information but also help enhance the model's predictive power.

Lastly, data set partitioning is a standard practice in machine learning.

Typically, the data is divided into training and testing sets, with the training set used for model training and the testing set for evaluating model performance. A reasonable data partition ensures the model's generalization capability on unseen data. Common partition ratios are 70% for the training set and 30% for the testing set, but specific ratios may be adjusted based on data volume and model requirements.

Through these steps, the quality of the data and the rationality of the model input are ensured, laying a solid foundation for building an accurate traffic congestion prediction model.

5. Application of Machine Learning Algorithms in Traffic Congestion Prediction

Traffic congestion prediction is a core function of Intelligent Transportation Systems, and machine learning algorithms play an essential role in this field. Machine learning algorithms can learn patterns from historical and real-time traffic data and predict future traffic conditions. The key to these algorithms is their ability to process large amounts of data and extract useful information, even when the data is highly complex and uncertain.

In the field of machine learning, various algorithms are suitable for traffic congestion prediction. Random Forest (RF) is an ensemble learning method that constructs multiple decision trees and combines their predictions to improve the model's accuracy and robustness. Support Vector Machine (SVM) can handle data classification and regression problems in high-dimensional spaces by finding the optimal boundary between data points. Neural Networks (NN), especially deep learning models, can automatically extract complex features and make predictions by simulating the way the human brain processes information.

When selecting algorithms, it is necessary to consider the characteristics of the data, the requirements of the prediction task, and the limitations of computational resources. For example, deep learning models may be more appropriate for data with obvious nonlinear characteristics, while random forests or support vector machines may be more efficient for smaller datasets. During the model construction process, it is necessary to define the inputs of the algorithm (i.e., feature variables) and the outputs (i.e., predicted traffic conditions), and choose the appropriate machine learning framework and tools for implementation.

Model training is a key step in the machine learning process, involving the use of training datasets to adjust model parameters to minimize prediction errors. During this process, it is important to monitor the model's performance to prevent overfitting or underfitting. Hyperparameter tuning is the process of adjusting the algorithm's hyperparameters to optimize model performance, with common methods including grid search, random search, and Bayesian optimization.

By training and tuning the model, a predictive model that performs well on the training set can be obtained. However, to ensure the model's generalization ability, it is necessary to validate it on an independent testing set. Test results will provide performance indicators such as model accuracy, precision, recall, and F1 score, which help assess the model's potential in practical applications.

In summary, the application of machine learning algorithms in traffic congestion prediction not only improves the accuracy of predictions but also provides data-driven decision support for traffic management and planning. As machine learning technology continues to advance, future traffic congestion prediction models will become more intelligent and efficient.

6. Construction of the Traffic Congestion

Prediction Model

Building an effective traffic congestion prediction model requires a comprehensive consideration of data characteristics, forecasting objectives, and algorithmic advantages. In this study, the model's architecture is designed following the principles of modularity and hierarchy, facilitating optimization and expansion. The input layer of the model receives data that has been preprocessed and feature-extracted, including multidimensional features such as timestamps, vehicle speeds, traffic volumes, and weather conditions. These features are transformed and combined through the feature layer to form new features indicative of traffic conditions.

In terms of the relationship between features and model performance, the research focuses on identifying which features have a key impact on predicting congestion. For instance, a decrease in vehicle speed and an increase in traffic volume are often precursors to congestion. Through correlation analysis, importance scoring, and feature selection algorithms, the most predictive feature set can be determined. Additionally, the model takes into account the dynamic nature of time series, using time windows to capture short-term trends in traffic flow.

Regarding the implementation details and algorithm description of the model, this study employs Recurrent Neural Networks (RNN) and Long Short-Term Memory networks (LSTM) from deep learning as the primary predictive tools. These network structures are particularly suited for processing time series data, capable of capturing long-term dependencies in traffic flow. LSTM units control the flow of information through three gates (input gate, forget gate, output gate), effectively addressing the issue of gradient vanishing in traditional RNNs.

During the model training process, Mean Squared Error (MSE) is used as the loss function to measure the discrepancy between predicted and actual values.

The optimization algorithm utilized in this study is the Adam algorithm, renowned for its efficiency in stochastic gradient descent, further enhanced by the adaptive learning properties of the RMSProp algorithm. To mitigate the risk of overfitting, the model training process incorporates regularization techniques, including dropout.

In the model evaluation phase, in addition to using traditional indicators such as accuracy, recall, and F1 score, special attention is also given to the model's ability to distinguish between different degrees of congestion (e.g., slight congestion, moderate congestion, severe congestion). Moreover, the real-time performance of the model is also an important consideration, as rapid changes in traffic conditions require the model to respond quickly.

Ultimately, the constructed traffic congestion prediction model demonstrated high predictive accuracy and good generalization capabilities on the test set. The model can analyze traffic data in real time and predict the occurrence of congestion in advance, providing a powerful decision support tool for traffic management departments.

7. Model Evaluation and Optimization

In Intelligent Transportation Systems, the performance evaluation of traffic congestion prediction models is a critical step to ensure their practicality and reliability. The choice of evaluation metrics directly affects the understanding and optimization direction of model performance. In this study, the evaluation metrics include accuracy, recall, and F1 score, which can comprehensively reflect the predictive capabilities of the model.

Accuracy measures the proportion of samples correctly predicted by the model to the total number of samples, which is a basic indicator for assessing the overall performance of the model. Recall, or true positive rate, focuses on the model's ability to capture positive samples, especially in traffic congestion prediction, where high recall means fewer missed congestion events. The F1 score is the harmonic mean of accuracy and recall, balancing between them to provide a comprehensive assessment of the model's precision and comprehensiveness.

To further verify the stability and generalization ability of the model, cross-validation methods were adopted. Cross-validation ensures the model's performance consistency across different data subsets by dividing the dataset into multiple subsets, with each subset taking turns as the test set while the remaining subsets serve as the training set. In this study, k-fold cross-validation was used, and the average values and standard deviations of the evaluation metrics were calculated to assess the model's stability and reliability.

In terms of model optimization, various strategies and methods were employed. First, the model structure was adjusted, including increasing or decreasing the number of network layers, adjusting the number of neurons, and introducing regularization techniques to prevent overfitting. Second, hyperparameters were optimized, such as learning rate, batch size, and the number of iterations, using techniques like grid search and random search to find the optimal hyperparameter combination. Additionally, different model ensemble techniques, such as bagging and boosting, were attempted to improve predictive accuracy.

Model optimization also included improvements in feature engineering, further refining features that are more indicative of predicting congestion through feature selection and transformation. At the same time, the data preprocessing process was optimized, such as adjusting standardization methods and strategies for handling missing values, to enhance the model's adaptability to data changes.

Ultimately, through a series of evaluation and optimization measures, the traffic congestion prediction model demonstrated high predictive accuracy and recall on the test set, and the F1 score was also significantly improved. The stability analysis results of the model showed that the fluctuations in the evaluation metrics from cross-validation were small, indicating that the model has good generalization ability and robustness. The application of these optimization strategies and methods has provided strong support for building an efficient and reliable traffic congestion prediction model.

8. Integration and Application in Actual Traffic Systems

Integrating the traffic congestion prediction model into actual traffic systems is a key step in realizing its application value. This process involves not only technical integration but also optimization and improvement of existing traffic management processes.

The application scenarios of the model are extensive, including urban traffic management centers, traffic signal control systems, and driver navigation services. In urban traffic management centers, the prediction model can provide real-time congestion prediction information for traffic dispatchers, helping them make more accurate traffic command decisions. Traffic signal control systems can adjust the timing of traffic lights according to the prediction results to optimize traffic flow and reduce congestion. For driver navigation services, the model can provide route planning suggestions based on predictions, helping drivers avoid congested areas and choose the best travel routes.

The technology and methods of system integration include the development of data interfaces, the architectural design of model deployment, and compatibility testing with existing systems. Data interfaces allow the model to exchange data with external systems, ensuring real-time information updates. Model deployment can use cloud computing platforms, leveraging their elastic computing resources to process large-scale traffic data. Compatibility testing ensures that the newly integrated model does not affect the stability and performance of existing systems.

Case studies and application effect analysis further verify the practical application value of the model. By deploying the model in specific cities or regional traffic systems and collecting traffic data before and after application, the model's impact on traffic conditions can be quantitatively evaluated. For example, by comparing traffic flow, average travel time, and the incidence of congestion events before and after the model deployment, the contribution of the model in reducing traffic delays and improving road usage efficiency can be assessed.

In practical applications, the integration and application of the model also face some challenges, such as the real-time nature of data, the stability of the system, and user adaptability to the new system. To address these challenges, it is necessary to establish corresponding technical support and maintenance teams to ensure the stable operation of the system and respond promptly to potential issues.

In summary, the integration and application of the traffic congestion prediction model in actual traffic systems not only improve the level of intelligent traffic management but also provide strong technical support for the sustainable development of urban traffic. With the continuous advancement and in-depth application of technology, future traffic systems will be more intelligent, efficient, and

user-friendly.

9. Conclusion and Future Work

This paper's research revolves around traffic data analysis and congestion prediction models in intelligent transportation systems. By comprehensively applying technologies such as data collection, preprocessing, feature extraction, machine learning algorithms, and model evaluation and optimization, an efficient and accurate traffic congestion prediction model has been constructed. The results of this study have not only enriched the theoretical field of intelligent transportation systems but also provided strong technical support for urban traffic management practices.

The summary of research findings indicates that the proposed model can accurately predict the occurrence of congestion based on real-time monitoring of traffic conditions, providing scientific and effective decision-making basis for traffic management departments. The application scenario analysis of the model in actual traffic flow shows that by integrating into urban traffic management centers, traffic signal control systems, and driver navigation services, the model can effectively improve traffic conditions and reduce the negative impact of congestion.

The significance and value of the model in actual traffic management are reflected in the following aspects: firstly, it improves the accuracy of traffic prediction, helping traffic management departments to formulate response strategies in advance; secondly, it optimizes traffic signal control and route planning, enhancing road usage efficiency; and lastly, it enhances the level of intelligentization of the traffic system, providing technical support for the construction of smart cities.

Although this study has achieved certain results, there is still room for improvement and expansion. Future research directions include further optimizing the model structure to improve the model's adaptability and generalization ability in different traffic environments; exploring more efficient algorithms and computing methods to meet the needs of real-time prediction; strengthening the interpretability of the model to help traffic managers better understand the prediction results; and researching how to more closely integrate the model with other components of the intelligent transportation system for more comprehensive traffic management.

In addition, potential pathways for model improvement include: training with larger-scale traffic data to improve the model's stability and reliability; introducing advanced technologies such as deep learning to automatically extract more complex traffic features; and considering the impact of socio-economic factors of the traffic system, such as holidays and major events, on traffic flow.

In conclusion, traffic data analysis and congestion prediction models in intelligent transportation systems have broad application prospects. With the continuous development and innovation of technology, future traffic systems will be more intelligent and efficient, better serving urban development and the lives of citizens.

References

Castro-Neto, M., Jeong, Y. S., Jeong, M. K., & Han, L. D. (2009). Online-SVR for Short-Term Traffic Flow Prediction under Typical and Atypical Traffic Conditions. *Expert Systems with Applications, 36*(3), 6164-6173.

Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1724-1734.

Daganzo, C. F. (2007). *Urban Gridlock: Theory, Causes, and Solutions*. Springer Science & Business Media.

Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation, 9*(8), 1735-1780.

Li, Y., & Zheng, H. (2012). Short-term Traffic Flow Prediction with Granger Causality Analysis in Consideration of Nearby Intersections Correlation. *Transportation Research Part C: Emerging Technologies, 20*(5), 1021-1036.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesne, E. (2011). Scikit-learn: Machine Learning in

Python. *Journal of Machine Learning Research, 12*, 2825-2830.

Rumelhart, D. E., Hinton, G. E., & McClelland, J. L. (1986). Learning Representations by Back-Propagating Errors. *Nature, 323*(6088), 533-536.

Smith, B. L., & Demetsky, M. J. (1997). Traffic Flow Forecasting: Comparison of Modeling Techniques. *Journal of Transportation Engineering, 123*(4), 261-266.

Van Lint, J. W. C., & Van Hinsbergen, C. P. J. M. (2005). Intelligent Transportation Systems: State of the Art and Future Prospects. *Transportation Research Part C: Emerging Technologies, 3*(1), 1-30.