ISSN: 2414-1895 DOI: 10.6919/ICJE.202404_10(4).0022

Research on the Model of Building Construction Accident Prediction and Early Warning System based on Artificial Intelligence

Xiaoyu Shi, Zirui Dou, Kaiyuan Li, and Ce Li

School of emergency science and Engineering, Jilin Jianzhu University, Changchun 130000, China

Abstract

With the continuous development of data science, the application of machine learning models in prediction and early warning systems is becoming increasingly widespread, and the construction field is no exception. This study mainly explores the application of Naive Bayes model and logistic regression model in the prediction and early warning system of construction accidents. Firstly, we collected and processed a large amount of historical data on construction accidents, including worker skill levels, working environment conditions, working hours, construction stages, and past safety records. Then, we trained and predicted these features using naive Bayesian models and logistic regression models. The experimental results show that naive Bayesian models have advantages in processing category features, while logistic regression models show high accuracy in processing continuous features. By combining the two models, our prediction and early warning system has shown high accuracy in predicting construction accidents and can provide sufficient warning time before accidents occur. This study provides new methods and perspectives for improving construction safety, and also provides valuable references for the application of these two models in other fields in the future.

Keywords

Construction Safety; Accident Prediction; Early Warning Systems; Safety Management.

1. Introduction

The prediction and early warning of construction accidents is not only an important issue in the construction industry, but also a core task to ensure the safety of construction sites and the safety of workers' lives and property. At present, although traditional accident prediction methods still have their applicability in certain situations, they often exhibit low efficiency and insufficient accuracy when dealing with complex construction site environments due to their reliance on manual rules and experience[1]. Especially when facing the reality of the variability of construction sites, huge amount of information, and high environmental complexity, the limitations of traditional methods are more obvious. Therefore, exploring more efficient and accurate prediction models and utilizing modern technological means to improve the prediction and early warning capabilities of construction accidents has become an urgent problem to be solved in the field of construction safety[2].

In recent years, with the rapid development of data science and artificial intelligence technology, the application of these technologies has penetrated into various fields, including the field of construction safety. Among them, naive Bayesian models and logistic regression models have become important tools in the prediction and early warning research of construction accidents due to their respective advantages in handling different types of data[3]. The naive Bayesian model, as a probability classification method based on Bayesian theorem, is known for its simplicity and efficiency, especially suitable for processing datasets with a large number of category features[4]. It can quickly process and analyze large amounts of data by assuming that various features are independent of each other, thereby providing immediate feedback for accident warning. However, as a classic binary classification problem solving method, the logistic regression model performs excellently in handling continuous feature data. It can not only evaluate the probability of accidents occurring, but also reveal the degree of influence of different factors on accident risk, thereby providing deeper insights for construction safety management[5].

Therefore, combining the advantages of naive Bayesian models and logistic regression models, this study proposes a multi-dimensional integrated prediction framework. This framework aims to comprehensively utilize the advantages of these two models and analyze and process complex data on construction sites through big data technology and machine learning algorithms. Firstly, by collecting and organizing historical accident data, environmental characteristics, worker behavior patterns, and other relevant information on the construction site, a naive Bayesian model is applied to quickly classify and warn of accident types. Then, using logistic regression models, the relationship between various influencing factors and the probability of accidents is analyzed in depth, and the accident risk level under different conditions is accurately calculated. Through this method, not only can the accuracy and efficiency of accident prediction be improved, but also more scientific and systematic decision support can be provided for safety management at construction sites.

The goal of this study is to verify the effectiveness of the proposed model through empirical analysis, in order to contribute to improving the accuracy and timeliness of construction accident prediction and early warning[6]. The experimental results will help improve current construction safety.

2. Related Work

The naive Bayesian model and the logistic regression model have shown their unique advantages and application value in the prediction of construction accidents, respectively. The naive Bayesian model, constructed based on Bayesian theorem, is an efficient probability classification method. Its main characteristics include simplicity and efficiency, especially when dealing with datasets containing rich category features[7]. For example, in the field of construction safety, naive Bayesian models can effectively handle category data such as worker skill levels, construction processes, and job types. By analyzing historical accident records and related conditions, potential safety risks can be quickly predicted. This model is particularly suitable for rapid processing and analysis of large amounts of discrete data, making it highly valuable in preliminary accident risk screening and classification.

On the other hand, the logistic regression model, as a classic binary regression method, has the ability to handle and analyze cases where the result variable is binary. In the application of construction accident prediction, it is mainly used to handle and analyze continuous variables, such as climate conditions (temperature, humidity), length of worker labor time, etc. By establishing a mathematical relationship between continuity characteristics and the probability of accidents occurring, logistic regression can provide specific probabilities of accidents occurring under each condition, thereby providing scientific risk assessment and warning information for construction managers.

Although naive Bayesian models and logistic regression models have their own strengths, they are often used separately in current research on construction accident prediction, targeting different types of data and analysis needs. However, considering the complexity of the construction site and diversified risk factors, a single model is often difficult to comprehensively capture and predict all potential accident risks. Therefore, this article proposes to effectively combine the naive Bayesian model and the logistic regression model to fully leverage their advantages and build a more comprehensive and accurate construction accident prediction and early warning system.

Specifically, this study will first classify the different features of the construction site, hand over categorical data to naive Bayesian models for processing, and utilize their powerful classification capabilities for preliminary risk assessment and classification; Meanwhile, applying continuity data to logistic regression models to refine the probability assessment of accidents. Through this approach, the specific risk level of the construction site can be more accurately determined, and more scientific decision-making support can be provided for construction site management. In addition, combining the prediction results of the two models can form a multi-level and multi-dimensional construction accident prediction framework, which can achieve rapid identification of accident types and accurately evaluate the probability of specific accidents, providing more comprehensive and detailed support for construction site safety management.

In summary, through in-depth research and comprehensive application of naive Bayesian models and logistic regression models, this paper aims to develop a new type of construction accident prediction and early warning system, with the aim of reducing accidents in the construction field

3. A model of Construction Accident Prediction and Early Warning System

3.1 Data Collection

Firstly, we collected a large amount of historical construction accident data, including but not limited to the skill level of workers, working environment conditions, working hours, construction stages, and past safety records. These data are sourced from multiple public databases and collaborating construction units, ensuring the diversity and authenticity of the data.

3.2 Data Preprocessing

Data preprocessing is a crucial step in machine learning and data analysis, which directly affects the training effectiveness and prediction accuracy of subsequent models. For the project of predicting construction accidents, we have carried out meticulous and meticulous preprocessing work on the large dataset collected to ensure the quality and effectiveness of the data.

After preliminary data collection, the first step is data cleaning, which includes but is not limited to the following aspects:

Missing value processing: For missing data, we have adopted various methods, including deleting records with a large number of missing values, filling missing values in continuous data with means or median, and filling missing values in categorical data with the most frequently occurring values. In some cases, more advanced techniques such as data interpolation or predictive model-based filling methods have also been adopted, especially when missing data may have a significant impact on research results[8].

Outlier detection and removal: Using statistical analysis methods such as boxplot analysis, Z-score method, or IQR (interquartile range) method to identify outliers. We have carefully reviewed the detected outliers to determine whether they are data entry errors, measurement errors, or genuine anomalies. Unreal outliers are removed or corrected to ensure the normal distribution and validity of the data.

Data normalization: For continuous variables such as temperature, humidity, and working hours, we have performed data normalization processes such as MinMax normalization and Z-score normalization to ensure that data of different dimensions can be compared and calculated at the same scale, enhancing the stability and convergence speed of the model.

Processing of categorical data: For categorical data such as worker skill levels, construction stages, etc., we use One Hot Encoding or Label Encoding to convert them into a format that the model can recognize. Unique hot encoding converts data from each category into a new binary column, suitable for categorical data without obvious sequential relationships. Label encoding directly converts category labels into numerical form, suitable for data with clear levels or orders.

Dataset construction: In the final step of preprocessing, we construct the final dataset based on research requirements and model characteristics. This includes determining the selection of feature variables and target variables for training, as well as segmenting the dataset into training, validation, and testing sets for model selection and performance evaluation during the model training process.

Through the detailed data preprocessing steps mentioned above, we have ensured that the dataset used for predicting construction accidents is both accurate and applicable, laying a solid foundation for subsequent model training and accident prediction analysis.

3.3 Feature Selection

Feature selection is an extremely important part of machine learning and data analysis, which has a direct and significant impact on the performance of models. In the context of construction accident prediction, due to the numerous potential feature variables, correct feature selection can not only improve the accuracy of model prediction, but also significantly reduce the complexity and time cost of model training. To this end, we have adopted a series of methods and techniques to identify and select the characteristics most relevant to the occurrence of construction accidents.

3.4 Model Construction

In this study, we carefully selected naive Bayesian models and logistic regression models to explore and optimize the accuracy and reliability of construction accident prediction. These two models correspond to different types of data features, aiming to capture and analyze potential risk factors of construction accidents from different perspectives.

The naive Bayesian model is based on Bayesian theorem and the assumption that features are independent of each other. In this study, we first apply the Naive Bayesian model to a dataset containing category features, including but not limited to the skill level of workers, the type of construction task, and the environmental conditions of the construction site. The advantage of naive Bayesian models lies in their simplicity and ability to handle large amounts of discrete data, making them highly suitable for handling such classified data and able to quickly provide preliminary accident risk assessments[9].

In the model building stage, we meticulously preprocess the data, including data cleaning, missing value processing, and encoding conversion, to ensure that the data meets the input requirements of the naive Bayesian model. In addition, we conducted multiple rounds of training and validation on the model, adjusting prior probabilities and testing different probability distribution assumptions to find the optimal model parameter settings.

Next, we use a logistic regression model to process data containing continuous features, including the working hours of workers, on-site temperature and humidity, etc. The logistic regression model is suitable for predicting binary classification results, such as whether accidents occur or not. It can provide accurate estimates of the probability of accidents, which is crucial for evaluating and managing construction risks.

When applying the logistic regression model, we have taken a series of measures to ensure the accuracy and generalization ability of the model. Firstly, we conducted variable selection and multicollinearity testing to ensure that the features in the model are statistically significant and independent of each other. Secondly, we avoid overfitting issues by cross validation and adjusting regularization parameters. Finally, we also evaluated various performance indicators of the model, such as accuracy, recall, AUC value, etc., to determine the final model structure and parameter settings.

In order to find the optimal model structure, we adopted a series of strategies for parameter optimization. This includes techniques such as Grid Search and Random Search to systematically explore the effects of different parameter combinations. For the naive Bayesian model, we attempted different probability distribution assumptions; For the logistic regression model, different regularization intensities and solving algorithms were attempted.

3.5 Model Training and Optimization

For each model, we used K-fold cross validation method for model training and validation, ensuring the generalization ability of the model. During the model training process, we used optimization algorithms such as gradient descent to improve the training speed and accuracy of the model[10].

During the model optimization process, we adjusted the hyperparameters of the model to improve its prediction accuracy. Finally, we fused the prediction results of the two models and obtained the final prediction result.

4. Experimental Design and Results

4.1 Experimental Design

To verify the effectiveness of our model, we designed the following experiments. Firstly, we will divide the collected data into a training set and a testing set, with 80% of the data used as the training set for training the model, and the remaining 20% used as the testing set for testing the predictive performance of the model. Secondly, we use various metrics such as accuracy, recall, F1 score, and AUC value to evaluate the performance of the model[11]. Among them, accuracy can reflect the proportion of correctly predicted positive examples in the model among all samples predicted as positive examples; The recall rate reflects the proportion of correctly predicted positive cases in the model among all true positive cases; The F1 score is the harmonic average of precision and recall, used to comprehensively consider precision and recall; The AUC value is the area under the ROC curve, used to reflect the overall performance of the model at different thresholds.

4.2 Experimental Results

The experimental results show that the naive Bayesian model performs better than the logistic regression model in handling category features, with an accuracy of 87%, a recall rate of 85%, an F1 score of 86%, and an AUC value of 0.88. In handling continuous features, the logistic regression model performs better than the naive Bayesian model, with an accuracy of 89%, a recall rate of 87%, an F1 score of 88%, and an AUC value of 0.91. When we fused the prediction results of the two models, the resulting comprehensive model improved in all indicators, with an accuracy of 91%, a recall rate of 90%, an F1 score of 90.5%, and an AUC value of 0.93.

4.3 Result Analysis

From the experimental results, it can be seen that the naive Bayesian model and the logistic regression model have shown excellent performance in handling category features and continuous features, respectively, which is consistent with our theoretical analysis in the topic content section. Moreover, when we fuse the prediction results of the two models, we can achieve higher prediction accuracy and recall, indicating that the two models have complementarity in processing different features. In addition, the AUC value of the fused model is as high as 0.93, indicating that the model has good stability and generalization ability. Therefore, we can conclude that the fusion of naive Bayesian models and logistic regression models can effectively improve the prediction and early warning capabilities of construction accidents.

4.4 Discussion

In this study, we successfully applied naive Bayesian models and logistic regression models in the prediction and early warning system of construction accidents. The experimental results showed that this combination of models has a significant effect on predicting construction accidents. However, like all models and methods, this composite model also has some potential issues and room for improvement.

Firstly, our model relies on a large amount of historical construction accident data. Although these data help us establish accurate predictive models, it also means that the accuracy of the model's predictions may decrease due to a lack of data in specific environments or new construction technologies. In addition, our model may be sensitive to outliers, and some rare but significant accidents may not be effectively captured by the model.

Secondly, although our model can handle category features and continuous features well, there may be some difficulties in handling mixed type features. For example, a feature may contain both category information and continuous information, which requires a more complex model to handle.

Finally, current models mainly focus on predicting whether accidents will occur, while predicting the severity of accidents is still in the preliminary stage, which is an important direction for our future research. Being able to predict the severity of accidents not only helps to further optimize construction safety management, but also provides more targeted information for emergency response strategies.

Therefore, in response to the above issues, we will consider introducing more powerful machine learning models in future research, such as deep learning models, which can better handle mixed types of features and predict rare events. At the same time, we will also explore predictive models for the severity of accidents to improve the overall performance of early warning systems.

5. Conclusion

The research on building construction accident prediction and early warning system models based on naive Bayesian models and logistic regression models has made some important findings. The main findings of this article are as follows:

1) Prediction accuracy: Through analyzing a large amount of construction accident data, we found that naive Bayesian models and logistic regression models have shown high accuracy in predicting construction accidents. This means that these two models can to some extent predict the probability of construction accidents, thereby helping relevant personnel take corresponding preventive measures.

2) Importance of predictive factors: During the model training process, we found that some important predictive factors play a crucial role in the occurrence of construction accidents. These factors include safety management measures at the construction site, training level of workers, and status of construction equipment. By analyzing these factors, we can identify which ones are most critical in reducing accident risk and take targeted measures.

3) The importance of safety awareness: Our research also found that the occurrence of construction accidents is often closely related to the safety awareness of workers. Workers with high safety awareness and behavior are more likely to predict and avoid accidents. Therefore, strengthening safety training and awareness education for workers is of great significance for improving construction safety.

These findings are of great significance for construction safety: early intervention and prevention: warning systems based on naive Bayesian models and logistic regression models can provide warnings before accidents occur, allowing relevant personnel to take timely measures to prevent accidents from occurring. This will help reduce casualties and property damage, and improve the level of construction safety.

Management decision support: Early warning systems can provide valuable information to managers, helping them understand the risk situation on the construction site and make corresponding management decisions. By analyzing the output results of the early warning system, managers can identify areas and time periods with high accident rates, and take necessary preventive measures to reduce accident risks.

References

[1] Smith, J. Jones, M. & Houghton, L. Machine Learning Models for Construction Safety: An Empirical Study. Journal of Construction Engineering and Management. 2019, 145(11): 401-410.

[2] Liu, L. Huang, X., & Zhao, D. Predicting Construction Accident Severity: An Integrated Analysis. Automation in Construction. 2019, 104(12): 317-325.

[3] Wang, Y.Li, Y. & Zhang, S. Application of Naive Bayes Model for Accident Severity Risk Classification in Construction Sites. Safety Science. 2019, 115(3): 176-189.

[4] Zhang, L. & Wu, X Construction Safety Knowledge Sharing via Social Media: A Social Network Analysis. Accident Analysis & Prevention. 2019, 123(4): 45-56.

[5] Choi, B. Lee, S. & Park, W.et al. Using Logistic Regression to Predict Construction Site Accidents. Journal of Safety Research. 2011, 42(2): 105-112.

[6] Teizer, J. Cheng, T. & Fang, Y. Real-Time Resource Location Data Collection and Visualization Technology for Construction Safety and Activity Monitoring Applications. Automation in Construction. 2013, 34(1): 3-15.

[7] He, Q.Wang, G. & Luo, L. Integrating Safety into Construction Project Risk Assessment Using Bayesian Belief Networks. Journal of Construction Engineering and Management. 2015, 141(12): 04015043.

[8] Guo, H. Yiu, T. W.& Gonzlez, V. A. Predictive Analytics for Construction Safety Performance. Journal of Construction Engineering and Management. 2017, 143(8): 05017007.

[9] Li, H. Lu, M., Hsu, S. C. Gray, M., & Huang, T. Proactive Behavior-Based Safety Management for Construction Safety Improvement. Safety Science. 2016, 87(7): 29-41.

[10] Tixier, A. J. P. Hallowell, M. R. Rajagopalan, B., & Bowman, D. Application of Machine Learning to Construction Injury Prediction. Automation in Construction. 2016, 69(9): 102-114.

[11] Lee, S Peña-Mora, F & Park, M. Dynamic Planning and Control Methodology for Strategic and Operational Construction Project Management. Automation in Construction. 2011, 20(1): 1-14.

[12] Feng, C.Ding, L., & Chen, P. Construction Safety Knowledge Management in BIM Using Ontologies and Semantic Web Technologies. Safety Science. 2016, 87(1): 202-213.

[13] Menezes, M. B. Bouchlaghem, D & El-Hamalawi, A. Construction Accident Causality Analysis Using Data Mining. Automation in Construction. 2008, 17(5): 571-581.