Predicting Crime Classification Using KDD Methodology

Syeda Nida Khader
San Jose University
September 30, 2024

Abstract

This research explores the use of the Knowledge Discovery in Databases (KDD) methodology to classify crime incidents based on attributes such as victim demographics, weapon used, and premises type. The dataset, sourced from the Los Angeles Police Department's Crime Data repository, includes crime reports from 2020 to the present. By following the five stages of the KDD process—Selection, Preprocessing, Transformation, Data Mining, and Interpretation—this study develops a Random Forest classification model to predict the type of crime (categorized as Part 1 or Part 2 offenses). The final model achieved a high accuracy and provided valuable insights into crime patterns across various regions.

Introduction

Crime data analysis plays a crucial role in enhancing public safety and assisting law enforcement agencies in crime prevention. With the advent of large-scale data collection, predictive modeling can be used to identify patterns and trends in crime, allowing for more efficient allocation of resources. The goal of this research is to develop a machine learning model that can classify crime types using the KDD methodology, a structured approach that guides the entire data mining process. By leveraging attributes such as victim age, sex, weapon type, and crime location, we aim to predict whether an incident falls under a Part 1 or Part 2 offense.

Methodology

The Knowledge Discovery in Databases (KDD) methodology consists of five phases: Selection, Preprocessing, Transformation, Data Mining, and Interpretation. Each phase contributes to building a comprehensive and accurate model for crime classification.

Selection

The dataset, "Crime Data from 2020 to Present," was sourced from the Los Angeles Police Department's open data portal. The data contains over 200,000 rows and 28 columns, covering various aspects of reported crime incidents. For this study, we selected relevant columns such as Vict Age, Vict Sex, Vict Descent, Premis Cd, and Weapon Used Cd to construct the feature set.

- Target Variable: Part 1-2 (Indicates whether the crime is classified as a serious Part 1 offense or a less severe Part 2 offense).
- Feature Variables: Victim demographics (Vict Age, Vict Sex, Vict Descent), crime circumstances (Premis Cd), and weapon type (Weapon Used Cd).

Preprocessing

The dataset was initially inspected for missing values and inconsistencies. Key preprocessing steps included:

- Handling Missing Data: Rows with missing values in critical columns such as Vict Age and Vict Sex were dropped to ensure data quality.
- Data Cleaning: Column names were standardized by removing leading and trailing spaces, and unnecessary columns were dropped to reduce dimensionality.
- Categorical Encoding: Categorical variables such as Vict Sex and Vict Descent were converted into numerical values using one-hot encoding.

Transformation

Feature engineering and transformation were applied to improve model performance:

- **One-Hot Encoding:** Categorical features were transformed into binary columns using one-hot encoding to make them suitable for the machine learning model.

- Feature Selection: Only the most relevant features (victim demographics, crime location, and weapon used) were retained based on correlation analysis.

Data Mining

The data mining phase involved training a Random Forest Classifier to predict the crime category (Part 1 or Part 2 offense). The Random Forest algorithm was chosen for its robustness and ability to handle categorical data effectively.

- **Model Training:** The dataset was split into training and testing sets (80% training, 20% testing) using stratified sampling to maintain class balance.
- **Model Performance:** The model was trained on the selected features and hyperparameters were tuned using Grid Search for optimal performance.

Interpretation

Model performance was evaluated using classification metrics such as Precision, Recall, F1-Score, and the Confusion Matrix. The final Random Forest model achieved high accuracy, indicating its effectiveness in classifying crime types.

- Precision: 0.87
- Recall: 0.84
- F1-Score: 0.85

The confusion matrix showed that the model was able to distinguish between Part 1 and Part 2 crimes with minimal misclassification, making it a reliable tool for crime type prediction.

Results and Discussion

The Random Forest model achieved an overall accuracy of 85%, with the most influential features being Vict Age, Vict Sex, and Weapon Used Cd. The analysis showed that certain types of crimes (e.g., assaults, robberies) were more likely to involve younger male victims, whereas Part 2 crimes (e.g., minor thefts) showed a higher likelihood of involving female victims. The strong performance of the model can be attributed to the structured approach provided by the KDD methodology, which ensured that each phase was thoroughly executed.

The visualization of the confusion matrix revealed that the model was more prone to misclassifying Part 2 crimes as Part 1, possibly due to the overlapping nature of features in certain cases. Future improvements could involve incorporating additional features such as time of day and socio-economic indicators to further refine the classification accuracy.

Conclusion

This study demonstrated the successful application of the KDD methodology to build a crime classification model using a real-world crime dataset. By systematically following each phase—Selection, Preprocessing, Transformation, Data Mining, and Interpretation—a robust and deployable model was created, achieving a high accuracy of 85%. The research provides valuable insights into crime patterns and offers a predictive tool that could be utilized by law enforcement agencies to prioritize resources more effectively. Future research could explore deep learning models and temporal analysis to further enhance the predictive power of the model.

References

1. Los Angeles Police Department Crime Data Repository. Retrieved from: LAPD Crime Data
2. Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1), 5-32.
3. Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From Data Mining to Knowledge Discovery: An Overview. *Advances in Knowledge Discovery and Data Mining*, 1-34.