STOR390 Midterm Project

Jonathan Zhao

2024-10-25

In today's fast developing society, there has been an increase in ideas on combating crime, many of which are utilizing the latest advances in technology and data collection. One of these ideas is predictive policing, which tries to predict future crime with data on past crimes. An example of predictive policing is Geolitica(previously called PredPol), which uses historical crime data to try to predict potential hotspots where crimes are likely to occur. However, a downfall of this method is that it can't be used during an active and immediate situation. In 'Crime Prediction For Better Response Using KNN, PCA, And Random Forest', Uchenna Akujuobi writes about how those methods don't help police officers prepare to respond to crime calls in real time.

The paper brings up two good examples to demonstrate this. The first example is the situation where a victim is hiding from dangerous people and can't make a 911 call without risking their life. This causes the victim to be unable to communicate the details of their situation with the police, and the only option left is for the victim to keep hiding and hope they are not found. The second example is when a police officer receives a crime alert with little information, and goes to the crime location not knowing what to expect. This could potentially put the police officer in a dangerous situation, or require the police officer to make quick impulse decisions which could lead to a negative outcome. In these situations, predictive policing is unable to be of any use since its main purpose is in predicting future crime, not impacting active crime. Developing a way to report and predict these crimes in real time could help police officers respond to crime calls accordingly and reach better outcomes.

This is where Akujuobi's research idea comes in. Instead of attempting to predict future crime using past data, Akujuobi's goal is to make real time predictions based on real-time information of a crime, to help police better deal with the situation. This model would entail creating a way for victims to report crimes without putting themselves in danger, such as a clicking a button on a mobile phone and a quick few words describing the situation. Next, the model would then use the information gathered to predict the type of crime being committed, and notify the police department of the possible crimes being committed. Using this information, police can make better decisions on how to respond, such as sending in specialists or additional reinforcements.

The main method that Akujuobi uses in this study to create this prediction system was to first predict the ward and district of the crime using K Nearest Neighbors and PCA, and then use multiple different algorithms to predict the crime that might have occurred in that area using features from a crime dataset. The research uses the Chicago Crime dataset, which includes crimes from 2001 to April 2015 when it was downloaded. This dataset has a total of 22 features, and the research uses seven of those features. These features are Location Description, Time, District, Ward, Longitude, Latitude, Primary Description(Target class) and Secondary Description(The last word). To deal with missing fields in some of the entries, any entries with missing fields were discarded to ensure the accuracy of the prediction result. The dataset was then divided into a training and testing partition, with a 75%/25% split.

Next, the prediction of the ward and district were made using the KNN classification method. For predicting the ward, a K value of 3 was used to achieve a prediction accuracy of 99.86%. For predicting the district, an accuracy of 99.97% was achieved from using a K value of 1. The reason that the ward and district were predicted and not taken from a simple internet search of Chicago's city database of district and wards were for the purpose of using the model out in the field. This enables the model to avoid potential issues such as bad internet connection or low speeds. After using KNN, PCA was applied on the features to reduce the dimension of the dataset and also map the dataset into a new space for maximum variance. The features with the highest 3 eigenvalues were taken and the rest were discarded to reduce dimension, which could help algorithms that did not scale well such as KNN. With the

reduced dimension and new space, Akujuobi then ran multiple simpler algorithms to try to predict the crime given the above features. These algorithms are K Nearest Neighbors(KNN), Artificial Neural Networks(ANN), Support Vector Machine(SVM), Decision Tree, Random Forest, and K-D tree algorithm of KNN.

For KNN, the Euclidean distance metric was used to determine the k nearest class labels. For ANN, multiple activation functions were attempted, but all took a long time to run and none gave any good result. SVM also had the same problem of long runtimes, and the model was shut off without completing after running for five days. Decision tree gave good results and was used in the evaluation of the method. For Random Forest, a different subset of the training data was used, using 2/3rds of the training data with replacement for training each tree. The remaining training data was for estimating error and variable importance. A total of one hundred weak classifiers were used in training the Random Forest model. Finally, the K-D tree algorithm of KNN was used to calculate the possibility of a predicted crime to be another crime, and was combined with Random Forest to catch the wrong predictions. This used a K value of 10.

To evaluate the accuracies of the different algorithms used, the algorithms were listed out based on highest accuracy. Additionally some of the algorithms were evaluated with and without the dimension reduction from using PCA. The highest accuracy was from using Random Forest at 88.99%. The KNN model performed better after applying PCA, while the decision tree model performed worse after applying PCA, but both were below Random Forest.

As it is important to see whether the predicted crime was accurate so that misinformation is not given to police officers, the possibility of being a different crime from the predicted crime was also calculated, using the K-D tree. This method catches all of the wrong predictions that Random Forest made, and lists the possible crimes it could have predicted. An interesting finding is that the actual crime was always among the possible crimes that were predicted using the model. The paper includes an example of where the predicted crime from Random Forest was battery, while the actual crime was robbery. The K-D tree showed that one of the possible crimes the model could have reached was robbery. This shows that there could

be some relationship between the predicted crime, the real crime, and the other possible crimes predicted. With this example, the robbery could have been committed with battery and assault. Thus, the study argues that there would be no conflict in actions to be taken by the police since if you are prepared for one scenario, you are prepared for the other scenario since they are related. Overall, the combination of Random Forest and K-D tree gives a good prediction of what is really going on at an active crime scene, and would help in deciding how police officers should respond.

The use of predictive policing introduces a few ethical concerns. A common normative concern that arises with crime prediction algorithms is the possibility of algorithmic bias. Because these algorithms are based on historical data, any societal or institutional biases from old policing practices could have an effect on the predictions. For example, if certain areas or demographics were disproportionately policed, then models based off of that data could also disproportionately predict for those areas or demographics. Akujuobi's work relies on Chicago's historical crime data from the years 2001 to 2015, so if there were any biases that disproportionately affected a certain racial or socioeconomic group, then Akujuobi's model may perpetuate such biases. Even though none of the features in the model explicitly contain any information about the crime's suspect, it is still possible for the algorithm to exhibit bias. This is known as discrimination by proxy, where an algorithm uses a different, innocent looking feature as a stand-in for a certain feature to reach a decision that should not be based on that feature. An example of this is the COMPAS algorithm, which is used to assess potential recidivism risk. Although it had no access to race, it was found that Black defendants were often predicted to be at a higher risk of recidivism than they actually were, and were also twice as likely as white defendants to be misclassified as a higher risk of violent recidivism (ProPublica, 2016). In this study, the model uses the location of the crime being committed as one of its prediction features. Many locations in Chicago may be demographically homogeneous, so using location as a predictive feature might cause the model to associate certain demographics to certain crimes, reinforcing stereotypes. If such an algorithm is used in the field, it may cause increased tensions between communities and law enforcement.

Another possible concern is the algorithm's ability to deal with false alarms. If the algorithm makes an incorrect prediction, it could have negative real world impacts, especially for the police who are relying on such predictions to approach a crime report. Akujuobi admits that the model is prone to false alarms and doesn't have the ability to detect them, meaning that it could lead to inappropriate police response. This could have severe impacts on the police force using the model, as it could escalate to over aggressive policing or even bring harm to those involved. False alarms can misallocate police resources, such as time, and even affect community trust. For example, if there is a false alarm in a neighborhood, and the model predicts it to be a severe crime, then police might send a large amount of resources there. This could cause tensions with that neighborhood, causing a sense of insecurity, as well as hurt the police force's ability to police in other locations.

Overall, Akujuobi's model could help in faster and better crime response for police officers in the field. With the predicted crime information, police would be better prepared to take on the crime call, enabling them to take on less unnecessary risks. The system could also help victims who are in dangerous situations to call for help without endangering themselves. However, this model is still not perfect, as the biases and issues that arise from using the model could cause potential issues. The field of combating crime with algorithmic analysis is still in its developing phase, and this research is a step in the right direction. As technology continues to advance and new models and methods are created, these issues should begin to diminish, and predictive policing will become more widely used.

References:

Akujuobi, Uchenna. Crime Prediction for Better Response Using KNN, PCA and Random Forest.

Larson,Julia, Jeff. "How We Analyzed the COMPAS Recidivism Algorithm." ProPublica, 23 May 2016, www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm. Accessed 26 Oct. 2024.

"Cogent | Blog | Predictive Policing Using Machine Learning (with Examples)." Www.cogentinfo.com, www.cogentinfo.com/resources/predictive-policing-using-machine-learning-with-examples.