Classification methods for Analyzing and Predicting Crime Trends in Chicago

Ricky Donnell Lindsey Jr

Georgia State University

College of Arts and Science

Department of Computer Science

CSC 4740 Data Mining

Abstract -Concern about rising crime rates in urban areas has continued to be a problem for modern developing society. Urban areas are active centers of commerce and a necessity for growing economies. By using supervised classification learners on standard information recorded by local police precincts, trends and correlation among crime data can better be understood and analyzed. By decision tress the predictability of data can be measured to tell how random a criminal act truly is. By using modified decision tress such ad boosted trees and bagged trees, the overall bias of data can be accessed and normalized to created models that better fit the current state of crime for a given area. Once identified proper governmental resources can be used to correct environmental problems that may be leading to increased crime rates.

1. Introduction

Concern about rising crime rates in urban areas has continued to be a problem for modern developing society. Urban areas are active centers of commerce and a necessity for growing economies. However, standard information on the geographical area's crimes is already collected in mass. This data usually contains information such as where the crime was committed along with the crime committed. By studying trends in this data problem areas within communities can be quickly identified, along with what crimes affect what areas, and what areas as are the most at risk.Crime is an escapable part of any modern society. Many individuals, groups, and organizations dedicate their lives to deterring an attempting to prevent crime. However, areas of highly concentrated crime are areas of high impoverishment. By identifying areas of high financially motivated crimes, such as theft, robbery and drug dealing, regions that are in the greatest need of the community and governmental intervention can be more precisely targeted.

1. Approach

All training and testing of data was performed using MathWorks’ Classification Learner Application within MATLAB R2017b. The first one hundred thousand entries from the Chicago Reported Crimes dataset was used for testing. Of the files twenty-two headers only ten headers were chosen. These headers were, the Illinois Uniform Crime Reporting codes (IUCR), the categorical description of the crime reported, the location of the crime reported (apartment, street, residency, etc.), was an arrest made, was it a domestic crime, the beat, the district, the ward, the community, and the FBI Code associated with the reported crime. The data was trained using fifty-fold cross validation with all ten features selected. Six classification learner models were used: Fine Tree Classifier, Medium Tree Classifier, Coarse Tree Classifier, Boost Trees Classifier, Bagged Tree Classifiers and RUSBoost Classifier. And All six classification learners used this same initial setup for the data.

The use of MATLAB’s Principal Component Analysis (PCA) was relied upon PCA allows for the reduction of redundant features in a model. This is possible do to the fact many groups of variables often move and trend in unison. By removing the redundantly moving variables it is thus easier to model and understand the data. The PCA was enabled to explain the variance of a given variable, by reporting what percentage of variance a given variable attributes to the model. All six models where built using a specified explained variance of 95%, or simply put, 95 % of the variance was tracked and attributed to the effects of a given feature. PCA only works on numerical data, not categorical. Therefore, only four features - Beat, District, Ward and Community – were numerical. Although the FBI codes are numerical values, the metadata for the file reports them as text, therefore categorical data.

1. Results

*Decision Trees*

Fine Tree

The Fine Tree model was the best preforming of all six models, with a 100% predictive accuracy. However, this does not mean it did not produce any wrong predictions. MATLAB measures only to a tenth of a percent accuracy, but that still means that with a fifty-fold cross validation on a sample set of one hundred thousand entries that less than twenty inaccurate predictions where made on average. This helps to cement the notion that the data is very unbiased, with very clear trends in crimes committed.

Medium Tree

The Medium Tree model performed the fourth best of all the models. The medium tree was trained to see if the sheer size of the data was affecting the apparent noisy ness of the data. However, with a predictive accuracy 86.2%, the Medium tree model helps illustrate that the bias of the data should be either closer to one of the extremes.

Coarse Tree

The Coarse tree model performed the worst of all six models. Having a predictive accuracy of only 66.5%, the Coarse tree model makes it safe to infer that the data is rather standardized with very few instances of irregular occurrences. This is very consistent with how many people would perceive the nature of crime. Crime usually occurs regularly in certain areas with particular crimes being more prevalent in a certain area.

*Ensemble Classifiers*

Boosted Trees

The Boosted Trees model preformed the third best with a 94.8% predictive accuracy. MATLAB boosted trees use a modification the Breiman method of gradient boosting. By using the optimization concept of steepest decent boosted tree can minimize overfitting controlling how far the data trends towards as particular value before correcting itself. The high predictive value of these model shows that there is minimal over fitting to the Chicago crime data. This helps to prove there is a regularity to the crime that can be modeled accurately.

Bagged Trees

The Bagged Tree model returned the second highest predictive accuracy of 99.9% of its predicted values correlating to accurate events in the test data set. The MATLAB bagged trees algorithm is a bootstrap aggregation using Brieman’s random forest algorithm. This methodology works most efficiently on trees than have a high amount of variance and low bias. From the high accuracy on this data it can be interpreted that reported crime in Chicago is variable in nature thus consisting of many crimes reported with no bias towards any rime being over reported.

RUSBoost Trees

The RUSBoost model performed the second worst of all the models, with a predictive accuracy of 68.8%. A RUSboost tree in MATLAB is designed for skewed data. It attempts to normalize the data. The low predictive accuracy of this model helps to strengthen the regularity of the crimes reported in the Chicago Crimes dataset. This furthers the notion that the data has minimal bias and an even distribution of the crimes reported.

PCA

The PCA was very insightful for understanding the redundant nature of this data. The model was built from only 6 of the initial ten features. Two of the four numerical features where deemed redundant. Since MATLAB doesn’t explicitly state which features are being suppressed, several training trails of the fine tree where run without PCA enabled with two of the 4 features suppressed until the best results were produced. These two features where the Ward and Community. The ward and community are essentially further specifications of locations within a district thus they are dependent upon the district. Since the community is the most specific location with the highest variance in its values. Therefore, the most valuable feature the community, thus rendering the district and ward to be redundant data that is not needed in building the model.

The PCA for all six models returned a variance per component of 62.5% for the Beat and a variance per component of 37.5% for the Community. This illustrates that the rate of reporting is more dependent on the Beat than the Community. Thus, the predictability of a crime varies more on the beat than the community.

1. Further Applications

This experimentation only scratches the surface of the possibilities of what can be done with this data. The understandability and workability of this data set were largely hampered by the hardware limitations the testing methodology was performed on. The experimentation with different learners served as an insight into the data that may be gleam by using different classification learners. However, there are a plethora of advancements that can and should be made for the betterment of research into this subject.

Hardware

As data continues to grow at an exponential rate, more advanced algorithms and machinery will be needed to compute these tasks. The live open data file for Chicago Crime Reporting had over forty-five million entries as of April 12, 2018. With a file size of only ~1 GB. The process of loading into hardware and processing the data was too daunting a task for a standard 6 cores 12 thread 16 Gb RAM Windows workstation. The system was limited to an approximate maximum of one hundred thousand entries.

The sheer number of entries and the number of calculations that need to be run also increase the run time of the models learning process. However, machine learning is parallelizable. The addition of computation oriented GPUs could greatly increase runtime efficiency. The use of cluster computing would also help in distributing the workload for more efficient run time and use of resources.

Software

The use of a more robust and efficient program language and API would be of greater deficiency in developing models to interpret data from for experimental use. Many cities across the nation track their crime rates. The establishment of a distributed network would allow for datamining to be performed on much larger sets of data. The use of an API such as Hadoop and Apache Spark, would allow for fast machine learning algorithms such as decisions trees to be performed on larger sets of data in a more robust language such as Java, Python or Scala.

Although graphical elements where not use in the initial model building process, visuals could better highlight at risk areas. Mined crime data could be overplayed d over maps to better identify troubled and potentially at-risk areas. Such feats may be possible by combining predictive models with already existing data to predict not only what are currently at risk but what areas may become at risk if current hos spots are properly dealt with.

1. Conclusion

All six classification learners are proof that crime in the City of Chicago is not statistically random and is predictable. By developing more accurate and robust models, communities can better handle crime problems. The use of classification learns is important in mining trends that better illustrate the nature of the data such as how variable the data is and how bias it may be.

1. References
2. J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques* [text]
3. Jason Brownlee, “Bagging and Random Forest Ensemble Algorithms for Machine Learning,” Aprill 22, 2016. [Online]. Available: <https://machinelearningmastery.com/bagging-and-random-forest-ensemble-algorithms-for-machine-learning/> [Accessed April 10]
4. Prashant Gupta, “Decision Trees in Machine Learning,” May 17, 2017. [Online]. Available: <https://towardsdatascience.com/decision-trees-in-machine-learning-641b9c4e8052> [Accessed April 10]
5. Sadanand Singh, “Understanding Boosted Trees Models,” Aug. 31, 2017. [Online]. Available: <https://sadanand-singh.github.io/posts/boostedtrees/> [Accessed April 10]
6. The MathWorks, Inc., “Documentation: Choose Classifier Options,” [Online]. Available: <https://www.mathworks.com/help/stats/choose-a-classifier.html> [Accessed April 10]
7. The MathWorks, Inc., “Principal Component Analysis (PCA),” [Online]. Available: <https://www.mathworks.com/help/stats/principal-component-analysis-pca.html> [Accessed April 10]
8. The MathWorks, Inc., “What Is Machine Learning: 3 things you need to know”. [Online]. Available: <https://www.mathworks.com/discovery/machine-learning.html> [Accessed April 10]
9. U. H. Atasever, E. Besdok, M. H. Kesikoglu, C. Ozkan, *THE USAGE OF RUSBOOST BOOSTING METHOD FOR CLASSIFICATION OFIMPERVIOUS SURFACES* [text]