State of California Crime Data: A General Framework for Data Mining

Hero Ozagho

February 15th, 2021

ABSTRACT

There had been an enormous increase in the crime in the recent past in the State of California. The concern is of great concern to national security and has significantly increased since the September 11, 2001 terrorist attacks at the Twin Tower in New York City, New York. However, information and technology overload hinders the effective analysis of criminal and terrorist activities.

Introduction

California's violent crime rate increased by 3.7 percent in 2016 to 444 per 100,000 residents. There have been other recent up-ticks in 2012 and 2015, but the statewide rate is still comparable to levels in the late 1960s. The state's violent crime rate increased dramatically from 1960 to 1980, from 236 to 888 violent crimes per 100,000 residents — a staggering 276 percent rise. After declining in the early 1980s, the rate rose to a peak of 1,104 in 1992. Since then, violent crime has declined substantially. California's violent crime rate is higher than the national rate of 386 and ranks 15th among all states. In 2016, 60 percent of reported violent crimes in California were aggravated assaults, 31 percent were robberies, 8 percent were rapes, and 1 percent were homicides.

Data Mining Techniques

Data mining is the process of looking at large banks of information to generate new information. Intuitively, you might think that data "mining" refers to the extraction of new data, but this is not the case; instead, data mining is about extrapolating patterns and new knowledge from the data you have already collected. Each of the following data mining techniques cater to a different business problem and provides a different insight. Knowing the type of business problem that you're trying to solve, will determine the type of data mining technique that will yield the best results. In today's digital world, we are surrounded with big data that is forecasted to grow 40 percent per year into the next decade.. The ironic fact is, we are drowning in data but starving for knowledge. Why? All this data creates noise which is difficult to mine – in essence we have generated a ton of amorphous data, but experiencing failing big data initiatives. The knowledge is deeply buried inside. If we do not have powerful tools or techniques to mine such data, it is impossible to gain any benefits from such data.

1 Classification

This analysis can be used to retrieve important and relevant information about data, and meta-data. It can also be used to classify different data in different classes. Classification is similar to clustering in a way that it also segments data records into different segments called classes. But unlike clustering, here the data analysts would have the knowledge of different classes or cluster. So, in classification analysis you would apply algorithms to decide how new data should be classified.A classic example of classification analysis would be our Outlook email. In Outlook, they use certain algorithms to characterize an email as legitimate or spam.

2 Outlier Detection

This refers to the observation for data items in a dataset that do not match an expected pattern or an expected behavior. Anomalies are also known as outliers, novelties, noise, deviations and exceptions. Often they provide critical and actionable information. An anomaly is an item that deviates considerably from the common average within a dataset or a combination of data. These types of items are statistically aloof as compared to the rest of the data and hence, it indicates that something out of the ordinary has happened and requires additional attention.This technique can be used in a variety of domains, such as intrusion detection, system health monitoring, fraud detection, fault detection, event detection in sensor networks, and detecting ecosystem disturbances. Analysts often remove the anomalous data from the dataset top discover results with an increased accuracy.

3 Clustering Analysis

The cluster is actually a collection of data objects; those objects are similar within the same cluster. That means the objects are similar to one another within the same group and they are rather different or they are dissimilar or unrelated to the objects in other groups or in other clusters. Clustering analysis is the process of discovering groups and clusters in the data in such a way that the degree of association between two objects is highest if they belong to the same group and lowest otherwise.A result of this analysis can be used to create customer profiling.

4 Regression Analysis

In statistical terms, a regression analysis is the process of identifying and analyzing the relationship among variables. It can help you understand the characteristic value of the dependent variable changes, if any one of the independent variables is varied. This means one variable is dependent on another, but it is not vice versa.It is generally used for prediction and forecasting.

5 Prediction

Prediction is one of the most valuable data mining techniques, since it's used to project the types of data you'll see in the future. In many cases, just recognizing and understanding historical trends is enough to chart a somewhat accurate prediction of what will happen in the future. For example, you might review consumers' credit histories and past purchases to predict whether they'll be a credit risk in the future.

6 Tracking Pattern

One of the most basic techniques in data mining is learning to recognize patterns in your data sets. This is usually a recognition of some aberration in your data happening at regular intervals, or an ebb and flow of a certain variable over time. For example, you might see that your sales of a certain product seem to spike just before the holidays, or notice that warmer weather drives more people to your website.

Discussion

The 2016 property crime rate of 2,545 per 100,000 residents is down 3.3 percent, about 3.5 percent above the 50-year low of 2,459 in 2014. Like violent crime, property crime increased dramatically between 1960 and 1980 from 3,140 per 100,000 residents in 1961 to a 50-year peak of 6,900 in 1980. But the property crime rate fell in the 1980s and '90s, and by 2011 it was down almost 63 percent. California's property crime rate is above the national rate of 2,450 and ranks 27th among all states. Of all reported property crimes in California in 2016, 64 percent were larceny thefts, 19 percent were burglaries, and 17 percent were auto thefts

Author contributions statement

Hero Ozagho, the author whose name is listed in this document certify that he has NO affiliations with or involvement in any organization or entity with any financial or non-financial interest (such as personal or professional relationships, affiliations, knowledge or beliefs) in the subject matter or materials discussed in this manuscript.