Machine Learning Engineer Nanodegree

Capstone Proposal

Niclas Geiger
June 9th, 2017

Proposal

Finding Resemblance to identify potential serial killers in homicide statistics

Domain Background

For solving murder cases there is always the need for information. There is a not that high change of solving homicide cases (61,5% in 2015)[1] which might also be attributed for finding suspects in cases where there is no personal relationship to the victim. These cases might often be linked to serial killer behaviour which rather depends on a pattern of victims instead of personal relationship.

Problem Statement

Finding and identifying potential suspects in homicide offenses has always been a manual task. There is a lot of investigation done by hand which is often tedious and time consuming. Many cases can be solved by looking for suspicious behaviour in the victim's relatives. If the killer is not personally related to the victim this task has to be approached differently. Serial Killer behaviour is often characterized in two ways: a) Having several victims[2] and b) showing some pattern in their behaviour. This Pattern might include: Location, tools used, special ways of committing the crime, attributes in the victim.

Datasets and Inputs

For investigation of this issue I am using the murder cases set from kaggle[3]. This dataset includes all homicide cases in the timespan between 1980 to 2014 in the United States. This dataset includes several attributes, which include information about the victim, the responsible agency, the location and date, the weapon used and the potential perpetrator.

---

[1] https://www.statista.com/statistics/194213/crime-clearance-rate-by-type-in-the-us/
[2] This is mandatory for being identified as a serial killer
[3] https://www.kaggle.com/murderaccountability/homicide-reports

Solution Statement

The Solution will be the understanding of which attributes have the highest chance of detecting serial killer behaviour. This will provide the possibility to reduce the state space to a much smaller one which in the end benefits in finding and identifying possibilities for serial killer behaviour in homicide reports.

Benchmark Model

My Benchmark model will be a set of several serial killer which I will identify by finding overlapping perpetrator attributes and researching according to these cases.

Evaluation Metrics

I will evaluate my model without the perpetrator attributes which would be a too strong identifier for serial killer behaviour. I will rather evaluate the relationship between cases depending on several attributes and finding resemblance or pattern in them. This knowledge can help to find serial killer behaviour in the future.

Project Design

First, I will try to identify several cases which might attribute to the same killer depending on their respective attributes. Then I will research if these cases can really be attributed to serial killer behaviour.
After that I will investigate the data set according to its attributes. I will use pattern and resemblance finding algorithms as e.g. KNN. Moreover, I will try to reduce the state space in the dataset to only reflect the fields which are really necessary for the analysis.
With the findings of the previous step I will try to identify which attributes are really the most precise in finding serial killer behaviour. This will help me develop a clustering method on a state space reduced dataset to identify serial killer behaviour also in the future. This method will then be benchmarked against. I will conclude my paper with researching[4] if this clustering gave me more serial killer cases as the benchmark model did.

[4] Researching means here to look for serial killer cases which might not be mentioned in the statistics so far or were not identified by looking at just the perpetrator attributes.