Learning Gun Violence Correlation

Mayur Patil
Dept. of Computer Science
University of California, Riverside
mpati005@ucr.edu

Abhay Singh
Dept. of Computer Science
University of California, Riverside
asing073@ucr.edu

Prateek Gambhir
Dept. of Computer Science
University of California, Riverside
pgamb001@ucr.edu

ABSTRACT

In the recent aftermath of the numerous high school shootings wherein many students were killed and several injured, firearm ownership has become a very polarizing topic. We collect data from a number of sources that maintain gun violence records and try to find insights from the data. We source our data from the Stanford Mass Shootings project, Mother Jones information website, and the gun violence archive. We study the data and associated medical studies and find the patterns in mass shootings, and the correlation between gun laws, literacy rate, average income, race, location with shootings. We also try to predict the shooting details using some known methods.

General Terms

Data Mining, Mass shootings, Behaviour analysis, Gun violence

Keywords

correlation, data mining, gun violence, mass shootings

1. INTRODUCTION

A gun violence incident where the number of casualties is beyond a certain number is deemed to be a mass shooting. The qualifying number differs from organization to organization and country to country. The US federal agencies keep the number at 4 casualties. For the scope of the project, we define mass shootings as gun violence incidents with 3 or more casualties. Due to a ban on gun violence research, official data for gun-violence and thus mass shootings has not been made public. Even federal agencies like the FBI and CDC cannot study or publish the data. However, due to efforts of a few independent organizations, we could create a dataset to study mass shootings.

In this course project, we intend to identify patterns in gun shootings, if any and derive the the correlation between the various factors. We do not wish to establish causality, however, we wish to provide evidence of correlation between several factors like race, literacy rate, income, location and others.

In this paper, we will discuss some of the established medical studies that will help us in identifying the correlations.

We have been able to identify several factors that may be correlated and use the data mining tool, Weka to perform several analysis techniques. Some of the classifiers and clusterers provide results with 95% plus accuracy.

2. DATA INFORMATION

2.1 Stanford Mass Shootings Project

The Stanford Mass Shootings of America (MSA) data project [5] began in 2012, in reaction to the Sandy Hook mass shooting. It is a curated set of spatial and temporal data about mass shootings in America, taken from online media sources. The MSA has stated that the data is not a comprehensive, longitudinal research project. The information collected for the Stanford MSA is limited to online resources. An initial intensive investigation was completed looking back over existing online reports to fill in the historic record going back to 1966. The project was abandoned in 2016.

Attributes in the dataset: location, number of fatalities, date, weapon used, motive, victims. Information about the shooter which includes race, age, mental illness, relation with the victims.

There are 335 data points.

2.2 Mother Jones Website

Mother Jones is a progressive American magazine that mainly writes on issues like new commentary, and investigative reporting on topics including politics, the environment, human rights, and culture. Its website [3] contains the dataset that spans from 1982 to 2018 is very limited compared to the MSA dataset, which implies they may have cherry picked the data to suit their story. The Mother jones dataset is used to complete the Stanford MSA dataset which was abandoned in 2016. Data-points beyond the 2016 were considered.

Attributes in the dataset: Location, date, victims, weapon used, motive, and whether or not weapon was obtained legally or not. Information about the shooter which includes age, sex, mental illness.

There are 99 data points.

2.3 Gun Violence Archive

Gun Violence Archive (GVA) formed in 2013 to provides free online[1] public access to accurate information about gun-related violence in the United States. GVA just like the Stanford MSA collects and checks for accuracy, comprehensive information about gun-related violence in the USA. It collects all incidents of gun related violence irrespective of the casualty count.

Attributes in the dataset: Location, casualties, number of injured people and news source.

There are 500 data points.

2.4 Mass Shootings Tracker

The mass shootings tracker[2] is an unfunded, crowd-sourced effort which includes data from Jan 2013 - Present. Users can report a mass shooting by submitting sources to the moderators of the subreddit that manages the database. The minimum sources cited for each incident is 2 or more. However, because it is not a verified source, we have used it for preliminary investigations only.

Attributes in the dataset: Date, name of shooter, number killed/wounded, city/state, and sources.

There are 1861 data points.

In addition to the above mentioned datasets, state literacy rates and income statistics were obtained from the official United States Government websites [6].

3. DATA PREPROCESSING

In this section, we want to discuss the need for creating a unified dataset from all the sources in order to increase the effectiveness of our analysis. Also, to be able to determine the relevant fields in terms whether they accurately represent the needs. We also wanted to ensure data remained consistent since it came from various sources.

3.1 Cleaning

To create a dataset which is a combination of several datasets, we had to ensure that we handle some of the attribute mismatches or missing attributes. We hence either eliminated several attributes or filled in values likes '0' or unknown. Any dirty data in terms of errors were eliminated. For example, we can never have negative number of victims. We have also eliminated duplicate records arising from different datasets.

3.2 Integration

A combined dataset was created with the base as the Stanford Mass shootings dataset. We have added data from Mother Jones datset and Gun violence archive. We have created a dataset that includes attributes like date, location, fatalities, injuries, mental health issues. We also ensure that no redundant data is present after integrating. For example, The recent Vegas shooting data was available in all three datasets with different titles and sources. However, important information like number of victims remained the same. Hence after careful examination, we use only one set of data pertaining to the Vegas shooting.

3.3 Reduction

We perform dimensionality reduction as the dataset involved had several attributes like summary which was a combined attribute holding the shooter information, victims information, location information and time information. However, we ignore this attribute as this may be helpful for news channels but not for data analysis. In the same way, the attribute title holding the information of the place of the incident and the number of victims was ignored. We also ignore some irrelevant attributes for our analysis like reported by attribute which has no impact on our analysis.

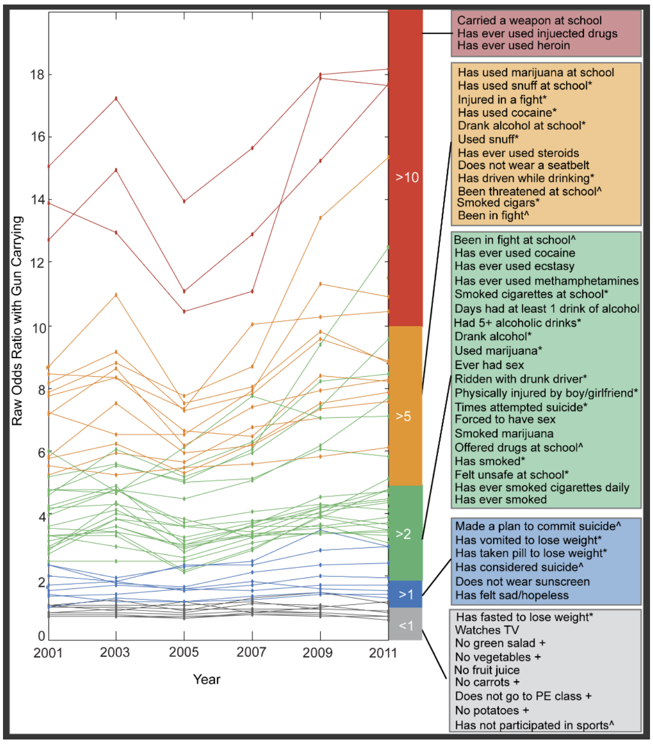

Figure 1: Gun violence association with risk behavior

3.4 Feature Generation

In some datasets, we do not have information about total victims that involves both fatalities and injuries while in some, it is available and in one of our analysis function, we use total victims of the incident and hence, we generated the attribute total number of victims where it was not available.

4. RELATED WORK

In this section, we want to discuss some of the previous work in the similar areas, some scientific studies related to the analysis we perform, some insight into the domain and some work on the tools and methods we use.

4.1 Correlating Gun possession with Gun Violence

This paper, "Gun Possession among American Youth: A Discovery-Based Approach to Understand Gun Violence" [14] tries to apply discovery-based computational methods to data from the Centers for Disease Control and Prevention's Youth Risk Behavior Surveillance System to understand and visualize the behavioral factors associated with gun possession.

Figure 1 shows us the different risk behaviour versus the odds ratio which is used to determine the associations [14]. The hierarchical clustering of risk behaviors(total 55) was done to visualize and behaviors that co-occur frequently, find

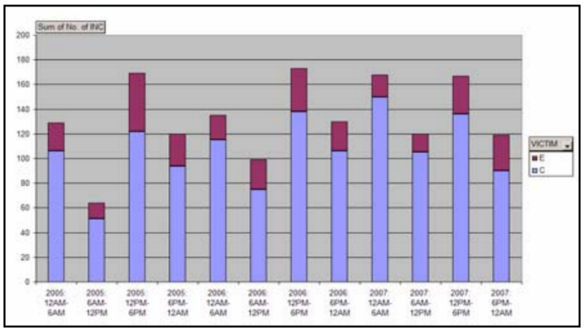

Figure 2: Assaults by 6 hour intervals

the association of which risk behavior that is most consistent with the gun possession.

The study is able to find six behavioural clusters namely, physical activity and nutrition; disordered eating, suicide and sexual violence; weapon carrying and physical safety; alcohol, marijuana and cigarette use; drug use on school property; overall drug use

- We are able to understand risk behaviors, beyond more commonly discussed indicators of poor mental health, that are associated with gun possession
- This work is based on the data, and does not provide strong basis for the associations with concrete medical evidences

4.2 Identifying patterns and trends in Violent Incidents

The paper, "Data Mining Customer and Employee-Related Subway Incidents" [7] talks about the preparation of dataset which originated from the New York City Metro Transit Authority (MTA) databases holding the information of date of assault, delay caused, locations, train times etc for the New York city subway system. The authors generate an attribute called as an event(incident type) with the data from the databases and try to predict the these event occurrences using the rule sets they generate and decision trees. This work was carried out using Weka, the data mining tool.

Figure 2, which shows the incidents that occur with a 6 hour interval[7], and we can see that most of the data within this work was used for creating visualization which in turn helped in understanding the incidents. The incidents are focused towards customers or employees and whether a weapon was used. This paper employs techniques like Apriori to determine rules. Some of such rules are

- With 100% confidence, a physical assault will occur against an employee
- With 100% confidence, incidents classified as robberies will not involve a weapon and a police investigation into the matter will result in a delay of the train

This paper highlights the needs of creating datasets matching our needs even if the data comes from different sources and importantly, it talks about the voilent incidents and what the metrics we should use to create a classification. It also gives insights into the effects of data preprocessing in correlation.
Pros and cons of the paper are

- We are able to understand the contents of datasets we must prepare for analysis of incidents involving violence
- We are able to understand the trees, cluster analysis can be performed on the data and time and location play a crucial role
- This work is old, from the year 2009 and most of the 'data cleansing' which is one of the major part of the paper can be done easily and efficiently using tools like OpenRefine[4]

4.3 Studying the impact of mental health, politics

The paper, "Mental Illness, Mass Shootings, and the Politics of American Firearms" [13], analyzes the most common arguments that are made aftermath of a gun related incident especially from a psychologist's perspective. This paper rather acts as a domain insight on the issue of mass shootings. The paper discusses 4 major points: First that mental illness causes gun-violence, second that some psychiatric diagnosis can predict gun crime, third that the shootings represent the deranged acts of mentally ill loners, and finally, that gun control will not prevent another mass shooting.

Talking about the first claim, the authors state although most mass shootings are perpetrated by individuals with at least some degree of mental illness most gun related crimes are not related to mental illness. Moreover, a number of most common psychiatric diagnosis like depressive anxiety and attention deficit disorders do not have any correlation to violence at all.

With respect to the idea of using psychiatric diagnosis to predict gun crime the authors state that although there have been gun confiscations in recent times based on mental health diagnosis. Historically psychiatrists have failed at being effective gatekeepers in this regard owing to the fact that psychiatric tools are observational in nature and not an extrapolative one. Another problem with the proposition is that psychiatric diagnosis and gun violence association has shifted over time: for instance soldiers with PTSD were not seen as violent, after the WWII ended but the idea of a crazy vet emerged with time.

The third claim was equally rebutted with citing historically racist backgrounds of the notion of a lonely mentally ill white male as around in the 50's there were a number of studies that claimed black culture was behind much of the then current mass shootings by black men. Raskin et al wrote, there are factors in negro culture that predisposes to more severe schizophrenic illness and thus violence eventually. The claim is also wrong statistically: even though there has been a rise in mass shootings by white lonely men lately, most mentally ill lonely white men do not commit such actions.

The final argument was also debunked by the authors, stating "Taking guns away from mentally ill will not eliminate gun violence" unless such efforts are linked to a broader impact on communities. Cities like Chicago-Illinois and Oakland-California have the strictest gun laws with respect to mentally ill individuals carrying guns, however they see very heavy gun violence on everyday basis. Oakland witnesses about 11 shootings daily.

To conclude the paper suggests among the most common arguments made in the wake of a gun related tragedy, most of them can not be any further from the truth. The only possible association that the authors validate is about mass shootings and mental illness.

4.4 Studying the Influences of Events

The paper, "Events and Controversies: Influences of a Shocking News Event on Information Seeking" [10] attempts to expand on how people navigate webpages on polarizing topics and it investigates whether online perusal reinforces preexisting viewpoints or depolarizes opinions and to what extent and whether the "filter bubble" phenomenon is real. It also determines, among other things, the effect of a shocking news event on browsing habits and viewpoints. The data is obtained from users' anonymized search through internet explorer's instrumentation between November and December 2012. It crawls Forums, advocacy groups, etc. to label domains as 4% factual, 2% highly balanced, 58% and 16% moderate and extreme gun rights, 18% and 2% moderate and extreme gun control respectively based on Normalized entropy of web page labels. The paper also examines existing theories such as whether people seek diversity or agreeable information before and after a shocking news event. The paper also examines overall transition patterns such as how the content of the current web page determines what the user will browse next. Finally, it also seeks to determine whether a shocking news event permanently shifts a user's viewpoint.

- It offers a model that can approximate inflection and trends on viewpoints which can further affect policy and ownership decisions.
- It also manages to introduce a mobility metric that can determine how viewpoints can change over time and after a shocking news event. It would be interesting to evaluate the model against events after the Sandy Hook shooting such as the Vegas and Florida shooting.
- By their own admission, it only includes data from the 29 million users of Internet Explorer and despite claims that it is a broad representation of the US population, it does not include other web browsers.
- Before the Sandy Hook shootings, the observed activity comes mainly from users who support gun rights. Also, their attempts to build content-centric classifiers did not achieve high accuracy, given the challenge of classifying controversial pages by their stance.
- The work can further be extended by inclusion of a larger and more diverse data set i.e browsers other than Internet Explorer. Attempt to build better content-centric classifiers.

4.5 Studying the impact of gun laws

The online article, "Predicting gun death rate from gun laws" by Jenny Listman uses a model that does a regression tree analysis of 18 years' worth of data from the CDC based on firearm death rates and gun laws by state[12]. It predicts death rates per 100,000 residents based off state gun laws in 10 categories.

It uses a codebook which includes each law, it's category and subcategory. However, it does not eliminate redundancy or correlation in the 133 individual laws and creates a list

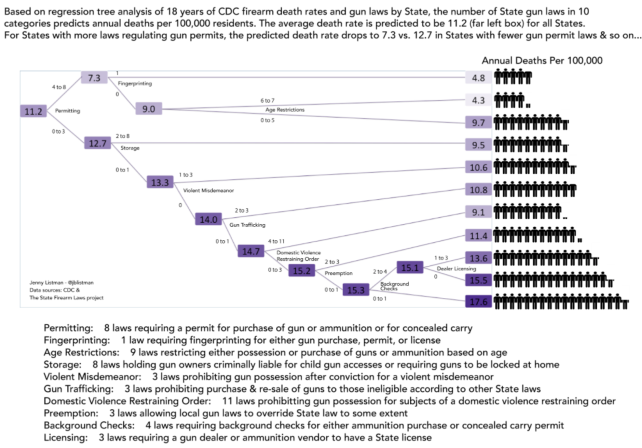

Figure 3: Listman's model Decision Tree

containing total number of laws based off subcategories in the codebook.

Data is split as 20:80 test and training sets. To account for correlation across states for a given year and correlation across years for a given state, the data set is split as odd or even years into 2 sets. Figure 3 shows the final decision tree.

- It offers a model to correlate the number of gun laws by state with firearm deaths using data obtained from CDC and Boston University's State Firearm Laws Project
- It also uses orthogonal data, since there are other factors at play in firearm related deaths such as unemployment rates and weighted population density per state.
- It still contains significant redundancy in the 133 individual laws and their subcategories such as grouping by type (dealer/client).
- Though it eschews factors resulting from orthogonal datasets , claiming an error rate of 1.76 deaths per 100,000, there are still significant number of factors from other data sets that it doesn't take into account.
- The work can be extended by inclusion of more orthogonal data to determine other factors affecting firearm deaths and the other extension could be elimination of redundancy related to gun law subcategories

4.6 Studying Counter Terrorism

The paper, "Data Mining for counter terrorism" [8] talks about how we can use data mining to predict and tackle terrorism acts. The paper proposes a method that uses online activities of individuals to find the âĂIJneedle in a haystack-âĂİ, the terrorist while also addresses privacy concerns that are usually raised with respect to any online surveillance. The author enumerates the types of terrorism and classifies them into two categories for the sake of analysis with data mining. A real time threat and non-real time threat. Non real-time threats are threats that do not have to be handled in real-time. That is, there are no timing constraints for these threats. Data is collected about a personâĂŹs behavior, where they have lived, their religion and ethnic origin, their relatives and associates, their travel records and segment critical and non critical data. Now the data is mined for any suspicious behavior like learning about flying a plane without the regard to learning to land a plane. Once done weâĂŹll have false positives and false negatives. These have to taken care by human inspection. A non real time threat can become a real time threat any point in time when a terrorist decides to attack. For the real time version data collection is a challenge as there are extremely few data points captured. However the paper proposes a methods to collect data from public surveillance cameras and sensors.

While addressing privacy concerns, author suggests that data collected can be partitioned into privacy sensitive and privacy insensitive data and stored. It will have to determined what data is private and to what extent. Author also claims that some data mining experts have argued that privacy enhanced data mining may be time consuming and may not be scalable.

4.7 Use of the data mining tool, Weka

The paper, "The WEKA Data Mining Software: An Update" [9] talks about the data mining tool, Weka and it's workbench which provides a comprehensive collection of machine learning algorithms and dara preprocessing tools. In the analysis of our Gun violence data, we wanted to perform specific tasks and based on that, we will briefly discuss sections from the paper on Weka. The explorer is a GUI that provides options such as Filters, Classifiers, Clusterers, Associations, and Attribute Selections. It provides a very good visualization tool in two dimensions.
Pros and cons of the paper are

- We get insights about how to effectively use the tool
- We get the visualizations that help in evaluation of the data. Certain decision can be based on such visualizations

5. DISCOVERING AND UNDERSTANDING PATTERNS

In this section, we briefly discuss some of the issues regarding the data or the topic and also some implementation details of the project.

5.1 Eyeballing the dataset

Studying any underlying pattern to separate gun violence incidents. Using the GVA dataset, we try to find if there is a common pattern of gun violence incidents. We assign each state a number. The numbers are assigned sequentially such that the difference in the numbers in neighboring states is small and states that are further apart will have higher numbers. The smaller states are assigned consecutive numbers. Bigger states have difference of 2 with their neighbors. Upon doing this the dataset can be viewed as a time series. Now the task is reduced to motif discovery.

5.2 Studying the visualizations

After we created a unified dataset, we used Weka to generate some visualizations to better understand how to go about trying to figure out the existing correlations between the attributes. How influential is the mental health of the shooter? Are shootings taking place only in specific locations? Do we have a reason to believe gender of the shooter is almost always male?

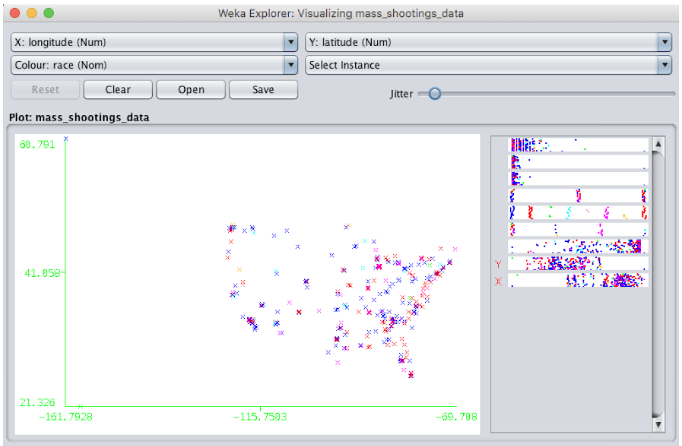

Figure 4: Locations of shootings in USA

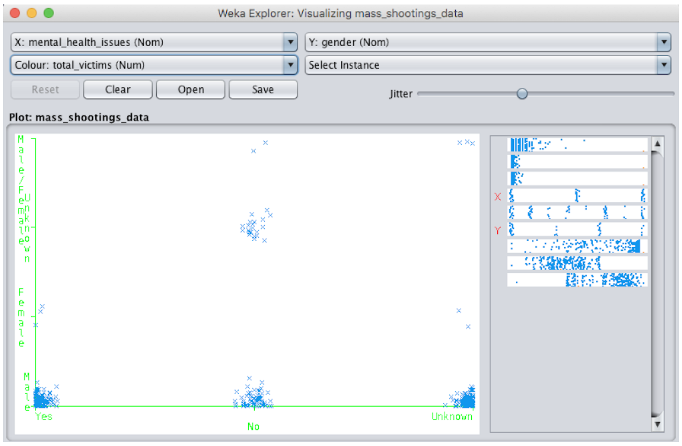

Figure 5: Mental health condition of shooters with their gender

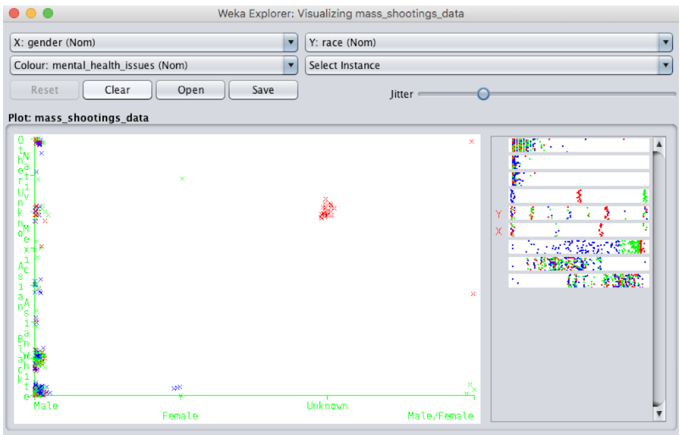

Figure 6: Shooters and their race

```
Cluster 0
(((((((((((((((((((((((((((((((-99.786168:0.04546,-95.824396:0.04546):0.01639,-


Time taken to build model (full training data) : 0.17 seconds

=== Model and evaluation on training set ===

Clustered Instances

0      319 (100%)
1        1 (  0%)
```

Figure 7: Hierarchical Clusterer applied to the data

Figure 4 shows the locations of the shootings in the USA. This is a helpful visualization as we can partition the shootings and try to identify if the shootings are location specific and try to investigate further why are there more shootings in these locations. What are the other factors we should be looking into based on this information.

Figure 5 shows the mental health condition of the shooters. Here we can see that there are male shooters have mental health conditions, do not have or is unknown, while female shooters without a mental health conditions are not present. However, there are few unknown mental health condition female shooters.

Figure 6 shows the gender of the shooters and their race. We can see that there are more White American shooters than Native American shooters. While there are almost the same number of Black American shooters as White American shooters.

5.3 Running Classifiers

We used Weka to run multiple classifiers on the data. The attribute selected was mental health issues which has three classes, yes, no and unknown and a percentage split of the data 66 % - 33% was used. The classifiers used and their accuracy of correctly classifying the instances is provided below

- Lazy KStar Classifier - 28%
- Naive Bayes MultiNominal Text Classifier - 44%
- Logistic Classifier - 91%
- Random Forest Classifier - 97%
- Bagging Classifier - 98%
- Classification via Regression Classifier - 98%
- Decision Table Classifier - 98%

5.4 Applying Clustering

We used Weka to apply multiple clustering algorithms on the data. All the models generated use the training data set. Some of the clustering techniques applied are

- Hierarchical Clusterer (Figure 7)
- Cobweb Clusterer (Figure 8)
- Kmeans Clusterer (Figure 9)

```
Number of merges: 159
Number of splits: 116
Number of clusters: 415

node 0 [320]
|   node 1 [91]
|   |   node 2 [43]
|   |   |   node 3 [19]
|   |   |   |   leaf 4 [1]
|   |   |   node 3 [19]
|   |   |   |   node 5 [5]
|   |   |   |   |   leaf 6 [1]
|   |   |   |   node 5 [5]
|   |   |   |   |   leaf 7 [1]
|   |   |   |   node 5 [5]
|   |   |   |   |   leaf 8 [1]
|   |   |   |   node 5 [5]
|   |   |   |   |   leaf 9 [1]
|   |   |   |   node 5 [5]
|   |   |   |   |   leaf 10 [1]
|   |   |   node 3 [19]
```

Figure 8: Cobweb Clusterer applied to the data

```
kMeans
======

Number of iterations: 8
Within cluster sum of squared errors: 321.9753506955729

Initial starting points (random):

Cluster 0: 5,0,5,Unknown,Black_American,Male,33.528287,-86.795504
Cluster 1: 0,4,4,No,Unknown,Unknown,26.640628,-81.872308

Missing values globally replaced with mean/mode

Final cluster centroids:
                                           Cluster#
Attribute                 Full Data               0               1
                            (320.0)         (212.0)         (108.0)
===================================================================
fatalities                    4.375          4.9811          3.1852
injuries                     6.1625          7.1651          4.1944
total_victims               10.1875         11.7547          7.1111
mental_health_issues        Unknown         Unknown              No
race                 White_American  White_American         Unknown
gender                         Male            Male            Male
latitude                    37.2251         37.4757         36.7331
longitude                  -94.4295        -97.5397        -88.3245
```

Figure 9: Kmeans Clusterer applied to the data

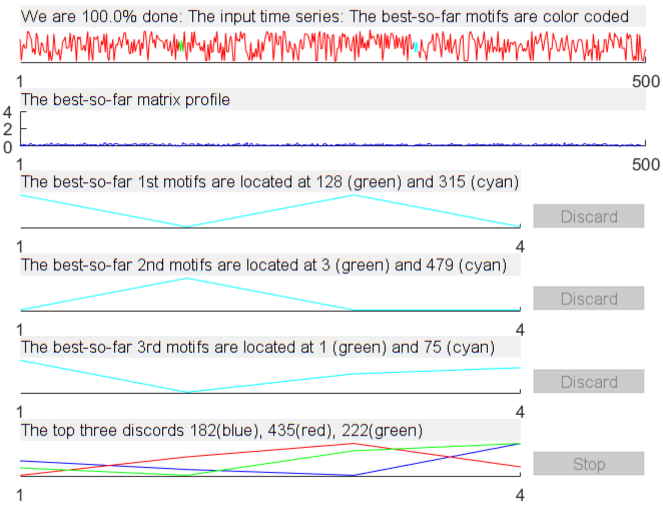

Figure 10: Generating Motif from the data using the Matrix profile tool

6. IMPLEMENTATION DETAILS

We perform some experiments on our data like Motif discovery for prediction, one class SVM classifier. We divide the work into the following tasks

- Spark SQL Tasks-
  - How many mass shootings per year
  - How many mass shootings per state
  - Mass shootings per Race/Gender/Age
  - How Many shooters had mental health issues and/or Social Issues.
- Motif discovery on Places. Is there a pattern in the shooting locations?
- Implement SVM to model the data and find outliers The below sections give a brief overview of our work

6.1 Finding Geographical Patterns in The shootings in USA

The Gun violence archive dataset was processed such that each neighboring state was given a relatively closer number while the states further apart were assigned numbers that had greater differences in their numbers. Sizes of the states were also taken into consideration. For instance, California was assigned 2 and Nevada was assigned 3. In another example Vermont was assigned 46.5 and New Hampshire was assigned 47. Once we had the state information converted into a time series, we could use the Matrix Profile tool[16] to find any motifs and discords. Some interesting observations can be made such as The matrix profile computed in window 2 suggests that although the MASS distance could be as high as 5(because of z normalization) the maximum MASS distance is much less than 1. The two motifs in each 3rd, 4th and the 5th window coincide with each other implying there are patterns that can be used to predict with some accuracy where the next shooting could happen.

- Figure10 shows the results obtained for query length 4

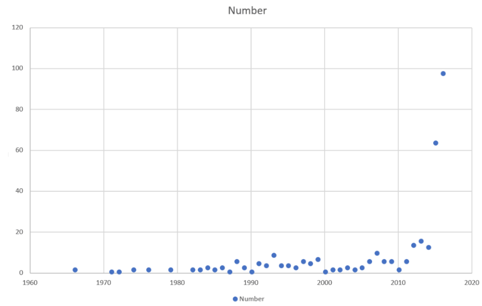

Figure 11: Yearly shootings incidents occurring throughout USA

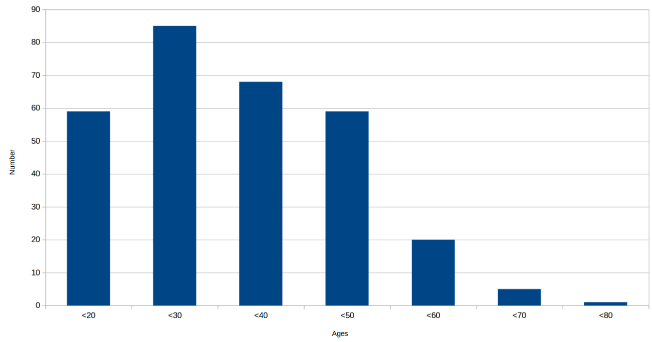

Figure 12: The figure shows different ages of shooters

- The matrix profile computed in window 2 suggests that although the MASS distance could be as high as 5(because of z normalization) the maximum MASS distance is much less than 1.
- The two motifs in each 3rd, 4th and the 5th window coincide with each other implying there are patterns that can be used to predict with some accuracy where the next shooting could happen.

6.2 One Class SVM Classifier

Since the gun owners data and mental health data is unavailable (no registry kept for gun ownership and health related data is not released to the public) the dataset built by amalgamation of Stanford MSA and Mother jones dataset

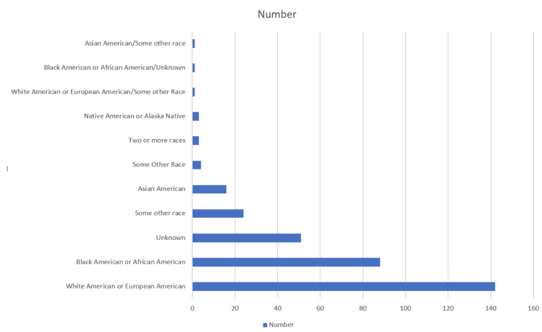

Figure 13: Shooters and their race throughout USA

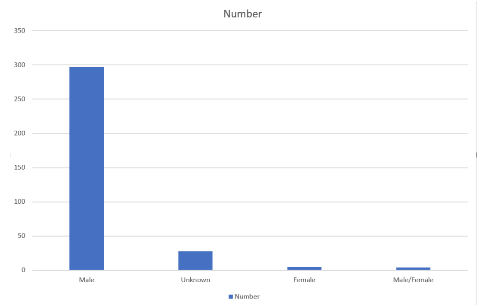

Figure 14: Shooters and their gender throughout USA

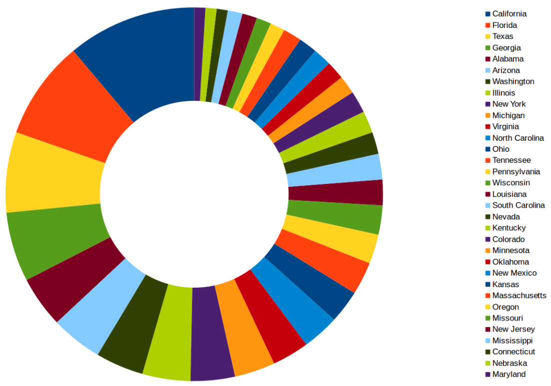

Figure 15: State-wise distribution of the shootings in USA

has only one class i.e. the shooters. This makes most models useless for us. However, as we saw above, most shooters had mental health and/or social issues we can build we can use these as a feature and do one class classification.

Most common one class classifier, the one class SVM [15] is also used to detect any novelty in the dataset. Hence it is also called a novelty detector. One-class SVM[11] is basically an unsupervised algorithm that learns a decision function for novelty detection: classifying new data as similar or different to the training set, thus classifying them. It uses a non-linear kernel for the decision boundary.

The data from Stanford MSA and Mother Jones Dataset was parsed and the dimensionality reduction was done by removing irrelevant features like lat-long, State, Name,City, etc. Only keeping the data attributes about the mental and social health, age, race and number of guns used. Please note the parameters that were unknown were assumed to be absent. For instance, if the mental issues of the shooter were not known, the shooter was assumed to mentally healthy.

For the implementation of the one class SVM we used, sklearn library in python. The code was written using jupyter and the other libraries that were used were numpy and pandas. The only features used were social and mental health issues. Off the total 336 data points, 37 data-points did not have information on either of the features and were considered mentally and socially healthy. 11 outlier data points were added to check if the model can detect outliers. The outlier points are supposed to emulate a healthy individual, i.e. mentally stable socially accepted person. Note total outliers thus become 48. 5 random 80-20 splits were done for training and testing datasets. The results were as follows

Measure	Output of the Model
PRECISION	0.8507462686567164
RECALL	1.0
ACCURACY	0.8507462686567164
F1 MEASURE	0.9193548387096774
AUC	0.67

7. RESULTS DISCUSSION

With the Motif discovery, we can see that

- There is an underlying pattern in the way the shooting happens. But in order to correctly predict where the next shooting may take place, the state to time series conversion pattern must be known correctly.
- Since the Matrix profile is always less than 1 we can use AB-Join Matrix profile[16] on a query window and the given history of gun violence to predict where the next shooting may take place.
- The prediction can be done with using the Bayes theorem, it has not been implemented for the scope of this paper. Extract the motif using the AB join Matrix Profile, with the corresponding window use Bayes theorem to find the probability of the next state where a shooting would happen. Report the state with the maximum probability.

With the One Class SVM Classifier, we can see that

- A one class svm model for the mass shooting dataset, with its only features being an individualâĂŹs social and mental well being (binary features) could detect outliers with 85% accuracy (detect normal people with 85% accuracy)
- The model can be improved by making the only two available features non binary. In other words, if we could factor in severity of mental illness and social issues of an individual we can more accurately detect a potential risk.

8. CONCLUSIONS

We are able to see that we can derive some correlations definitely exist between gender of the shooter where we can see shooters are mostly identified as male (98%). We can also see that without the outlier data points or the incomplete data points, we can say shooters are mostly identified as White American(32%) and Black American(36%). Also, we can see that the shootings mostly occur on the coastal region of USA(California being the majority. We are also able to see that we can predict some instances accurately with 85% using one class SVM classifier for Outliers.

9. REFERENCES

[1] Gun violence archive. *Online - http://www.gunviolencearchive.org/mass-shooting.*

[2] Mass shootings tracker. *Online.*

[3] Mother jones : Us mass shootings, 1982-2018: Data from mother jones investigation. *Online - https://www.motherjones.com/politics/2012/12/mass-shootings-mother-jones-full-data/.*

[4] Open refine - a free, open source, powerful tool for working with messy data. *Online*.

[5] Stanford libraries : Mass shootings in america. *Online - https://github.com/StanfordGeospatialCenter/MSA*.

[6] Us government websites. *Online collection*.

[7] J. J. David Budet, Mariel Castro. Data mining customer and employee-related subway incidents. *Proceedings of Student-Faculty Research Day, CSIS, Pace University, May 8th, 2009*.

[8] M. Derosa. *Data Mining and Data Analysis for Counterterrorism*. 2004.

[9] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. The weka data mining software: An update. *SIGKDD Explor. Newsl.*, 11(1):10–18, Nov. 2009.

[10] D. Koutra, P. N. Bennett, and E. Horvitz. Events and controversies: Influences of a shocking news event on information seeking. *CoRR*, abs/1405.1486, 2014.

[11] K.-L. Li, H.-K. Huang, S.-F. Tian, and W. Xu. Improving one-class svm for anomaly detection. *Proceedings of the 2003 International Conference on Machine Learning and Cybernetics (IEEE Cat. No.03EX693), 2003*, 2003.

[12] J. Listman. Predicting gun death rate from gun laws. *Online*.

[13] J. M. Metzl and K. T. MacLeish. Mental illness, mass shootings, and the politics of american firearms. *American Journal of Public Health*, 105(2), 240âĂŞ249., 2015.

[14] K. V. Ruggles and S. Rajan. Gun possession among american youth: A discovery-based approach to understand gun violence. *PLOS ONE*, 9(11):1–12, 11 2014.

[15] J. W. Y. Wang and A. Miner. Anomaly intrusion detection using one class svm. *Proceedings from the Fifth Annual IEEE SMC Information Assurance Workshop, 2004.*, 2014.

[16] C.-C. M. Yeh, Y. Zhu, L. Ulanova, N. Begum, Y. Ding, A. Dau, D. Silva, A. Mueen, and E. Keogh. Matrix profile i: All pairs similarity joins for time series: A unifying view that includes motifs, discords and shapelets. *IEEE 16th International Conference on Data Mining (ICDM), Barcelona, 2016*, 2016.