CS685: Group 10
Studying Crime Incidents for Safety Analysis

Arbaz Khan
arbazk@iitk.ac.in
Dept. of CSE

Mayank Dang
mayankd@iitk.ac.in
Dept. of CSE

Nikhil Pangarkar
nikpan@iitk.ac.in
Dept. of CSE

Indian Institute of Technology, Kanpur

Final project report
November 17, 2012

Abstract

San Francisco State Government has revealed all the crime incidents of the past ten years of San Francisco. Broadly, the data contained information in the form of dates of the incidents, the type of the crime and the exact location of the incident reported. We wish to analyse the correlation in the various attributes and try discover some patterns which could help the concerned authorities and the decision takers to formulate strategies to try improve the safety of the citizens. We would try to predict crime-vulnerable-areas dependent on the time using the available data. If successful, this could help aniticipate the crimes and realise a dream of crime-free state. The data consists of records of over hundred thousand reported incidents with each record consisting of attributes such category of crime, location, time, etc.

Introduction

In present world scenario, where major cities are prone to crime in great numbers, the crime analysts have to carry out a rigorous task of studying the crime statistics and assisting the authorities to cause arrests and crime prevention. However with the vast expanse of data and the lack of awareness about the locations of a city makes it even difficult for human to pursue the analysis tasks. There may well be some inherent patterns in the reported incidents which get missed by ordinary human perception. There are various machine learning and data mining tools and techniques which study data and carry out fruitful results of classifications and prediction on a data well analogous to a crime database. The benefits of such an analysis can be huge, if we think about the full potential of the results and findings. Say, for example, if using the patterns one could predict that there could be sale of drugs in the southern district of the city at the end of this week, then this information can be utilized to highten the security of that area.

Realizing the potential of a well maintained crime database, the police authorities around the world have taken steps to maintain quality data of the crimes reported. San Francisco State Government sits on the top of such intelligent institutions who have openly published their data to the world to either use it for one's personal statistical analysis or may help the state by notifying about certain patterns. The data from the past ten years is as voluminous as 100,000 records with as many as ten different attributes defining a record. Its not only inefficient but daunting task to carry out such an analysis by the police analysts. One requires a machine to recognise the patterns in the data accurately using a predefined metric. What we do here is no less than the prescribed needs.

Although there could be many types of analysis possible on a crime database including entity reconciliation for identifying an individual behind several crimes or clustering for identifying crime groups, the data available restricts us to geographical location and time related analysis. We exploit the fact here that crime is not randomly distributed and there maybe some latent factors yet unexplored, which do include an inherent geographical nature. Further accentuating the contribution of location to a criminal activity, there is a theory which states that crimes are related to offender's personal activity space and that the offenders are restricted to move around a certain place.

Such stands the role of location in crimes. Not only on the locations, crimes also depend upon time of the day and also upon the type of the crime themselves.We would try to discover the patterns these crime incidents follow, if any, and how could they be used to reduce criminal activities. This ability of predicting a crime before its occurence or getting some knowledge about its location can help the law enforcements to organise the patrol officials and help avoid the crimes or at the least reduce the response times to deal with such acts.

Problem Statement

Data

The data available to us contains about 130,000 records with each record being defined by a set of attributes. The attributes can be described as below:

- Incident Number;
- Category of the Crime;
- Crime Description;
- Day,
- Date and
- Time of the reported crime;
- Police District;
- Resolution status;
- Address;
- Latitude;
- Longitude

Objectives

The objectives that we set ourselves with the data available to us include identifying useful frequent patterns in the crime incidents and a framework to help predict hot spots(crime vulnerable locations) in a particularly defined scenario (e.g. time of the day, period of the year).

Related Work

There has been a lot of research done in the areas of crime prediction or predictive crime mapping [1] [2] [3]. Although there are various other famous areas for crime prediction which aim at estimating the dangerousness of a criminal, or identifying the individual behind several incidents and also identifying groups within the criminals. However, predictive crime mapping targets at predicting locations of crime. There have been tools, systems and software built to reduce crime using statistical analysis. Most notable of these is CRUSH, built by IBM under the collaboration of University of Memphis for the state of Memphis. It aims at predicting the location of future crimes using statistics. There have been various news about police departments having employed various techniques to be a step ahead of the criminals and trying to predict crimes before they happen. LAPD was reported to have done so recently [4]. Work is done on the same basis of identifying hot spots based on the incidents reported in the past. The existing prediction models have been analogously applied on crime databases as well. For example, Santa Cruz Police used earthquake prediction models for the purpose of forecasting crimes in the local regions. Although enough amount of activity has been already done in this field, the data of San Francisco, even being publicly available has had only statistical analysis and no predictive models have been proposed for it.

Approach

Data Reduction

Attribute subset selection requires selecting only the important attributes of the records and elimintation of those which provide redundant information. The attributes *Latitude*, *Longitude*, *Address of the Street* contained redundant information on the location of the crime incident and hence were transformed to one single attribute.

Incident Number was dropped as it just served the purpose of primary key in the original database.

Data Discretization

The area of San Francisco was divided into 900 bins and each location was represented by the rectangle in which it

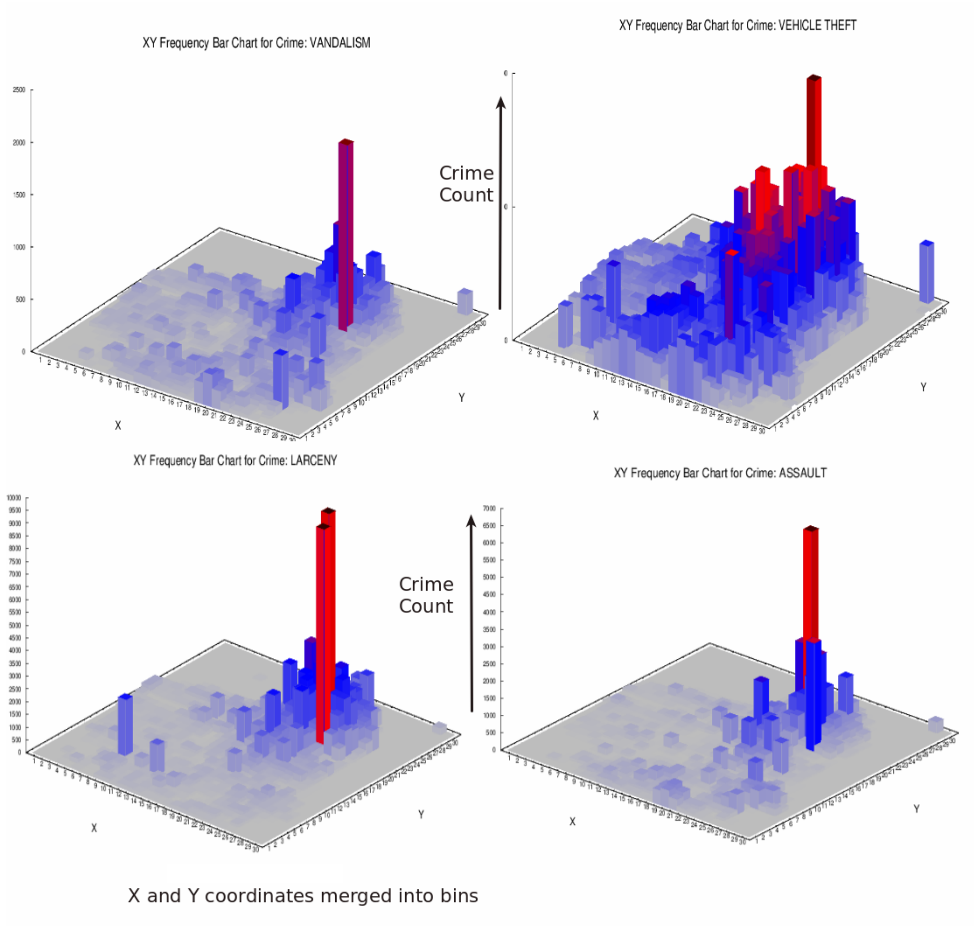

Figure 1: 3D plot showing the spatial distribution of reported crime incidents

contained. The rationale behind the number 900 was that there were 19000 unique street addresses. Considering the fact that the size of the data was around 100,000 such a large number of partitions would have led to sparse distribution. Also division according to Police Department districts led to only 10 partitions of the locations in San Francisco which would have resulted in loss of information due to under-fitting.

This discretization, however, resulted in the creation of grid blocks having no crime record in the given data. Such blocks had to be pruned as these areas are of no importance to the authorities.

Time was discretized into intervals of an hour each.

Graphical Analysis

To get an insight into the data and the location wise distribution of the reported incidents we plotted some 3D plots. The graphs corresponding to the most frequent crimes in the city are shown in the Figure 1.

Knowledge Discovery

To discover the inherent patterns in the crime records, we tried finding strong association rules among the data attributes. For the same purpose, we analysed the correlation among the attributes using lift as a measure. Although a very primitive approach it helped in discovering some patterns which are summarised in the results section.

Crime Prediction

Predictive analytics is a technique that extracts information using Statistics and Data Mining and uses it to predict future events. For predicting crime, one needs to look upon the factors that influence the occurence of crime events at a particular location at a given time. Here, we give a generalized approach for recognising temporal and spatial crime patterns and using them for crime prediction. We are using a classification-based model that takes into account past crimes at a given location with varying time dimensions/resolutions along with few other parameters depending upon the nature of prediction.

Taking into account the utility of crime prediction system, we worked upon three types of predictions. The approach behind each of these is mentioned in the following subsections. The crux of the approach in the different prediction types stays broadly the same and relies upon the spatial and temporal factors associated with crime events.

Predicting crimes at a given time of a day

To utilize the temporal factors affecting crime at a particular location, we take into account the crime history of that location. Here, we assume that the number of crime instances at a particular time of day in a month depends upon the instances at the same time in previous months. For example, we assume that the crime instances taking place at $4pm$ in the month of April depend on the crimes taking place at $4pm$ in the previous months of March, February, January and so on.

To take into account the spatial factors, we used the average crime count of the neighbourhood at the same time instance. Crime at a given location is significantly correlated with the crime in the neighbouring areas [5], [6]. Such an approach is also supported in other works aimed at predicting crime, such as [3]. This also reduces the error introduced due to discretization, if any.

Henceforth, the feature-set is constructed using the spatial and temporal factors described above. We consider a total of T+2 attributes, where T attributes denote the crime instances of the previous T months. The other attribute denotes the average crime count of the neighbourhood for the past T months and the hour under consideration.

Grid_no	Hour	Year_month	last month	second last month	third last month	average neighborhood	this month
620	21	200808	28	41	45	40.875	25
620	21	200809	25	28	41	40	31
620	21	200810	31	25	28	36.5	62
620	21	200811	62	31	25	31.875	44
620	21	200812	44	62	31	30.5	54
620	21	200301	0	0	0	0	30
620	21	200302	30	0	0	6.5	19
620	21	200303	19	30	0	13.625	24
620	21	200304	24	19	30	23.5	24
620	21	200305	24	24	19	26.875	26
620	21	200306	26	24	24	31.125	20
620	21	200307	20	26	24	31.25	18
620	21	200308	18	20	26	29.625	39
620	21	200309	39	18	20	28.875	27
620	21	200310	27	39	18	32.375	32
620	21	200311	32	27	39	31.75	18
620	21	200312	18	32	27	30.75	24
620	21	200901	54	44	62	28.75	20

Figure 2: DataSet- Predicting crimes at a given time of a day

Predicting crimes at a given day of the week

The approach here is similar to the one while trying to predict crimes at a given time of the day. We assume that the possibility of crime taking place at a particular day of the week in a month depends upon the crime instances at the same day in previous months. For example, we assume that crime instances taking place at $Wednesday$ in the month of April depend on the crimes taking place at $Wednesdays$ of previous months of March, February, January and so on. However, here we only predict whether crime takes place at Wednesdays in April or not.

The average crime count of the neighbourhood at the same day of the month is considered to take into account the spatial dependence of crime on the neighbouring areas.

Hence, the feature-set is constructed using spatial and temporal factors described above. We consider a total of T+1 attributes, where T attributes denote the crime instances of the previous T months. The other attribute denotes the average crime count of the neighbourhood for the past T months. However, we need to take care of the sparsity of the data of crime instances at a particular day of the week in a month. There do occur a relatively greater number of non-crime events while creating the dataset for the *day of week* crime prediction analysis. Prediction of crime not taking place in such a dataset is an unworthy achievement. We need to make the classifiers focus more upon the prediction of crime events. We use the balancing technique provided in [3] to increase the weight of crime events over non crime events. The weight factor to be added is the ratio of number of non crime events and the number of crime events.

month7	month6	month5	month4	month3	month2	month1	nbr_total7	this_month
55	77	68	44	41	86	61	1545	0
77	68	44	41	86	61	63	1519	1
68	44	41	86	61	63	85	1571	0
44	41	86	61	63	85	76	1546	0
41	86	61	63	85	76	53	1535	1
86	61	63	85	76	53	85	1570	1
61	63	85	76	53	85	91	1514	1
63	85	76	53	85	91	100	1523	1
85	76	53	85	91	100	95	1589	0
76	53	85	91	100	95	65	1526	1
53	85	91	100	95	65	88	1530	0
85	91	100	95	65	88	57	1495	0
91	100	95	65	88	57	77	1428	1
100	95	65	88	57	77	95	1458	0
95	65	88	57	77	95	58	1455	1

Figure 3: DataSet- Predicting crimes at a given day of the week

Time series analysis for predicting crime intensities at a given day

In the previous sections we have developed models to predict time pertaining to intensities in a given month. Here, we aim to take the predictions at a higher resolution. A prediction for crime intensive zone at a particular day would be of more significant interest to the police authorities as it would require immediate action. Violent crimes are the ones that require daily vigilance. So we consider these crimes as our focus of interest:

LARCENY/THEFT,ASSAULT, BURGLARY, KIDNAPPING, SUICIDE,ARSON, DRUG/NARCOTIC, VANDALISM, SEX OFFENSES- FORCIBLE, WEAPON LAWS,DISORDERLY CONDUCT

Next, we go about clustering the city regions so that the good enough daily data is available. The aim is to come up with a prediction of crime intensities in certain broad areas of the city on a daily basis. Each cluster is then seperately regressed upon to cope up with their respective average intensities of violent crimes. Subsequent sections elaborate the approach.

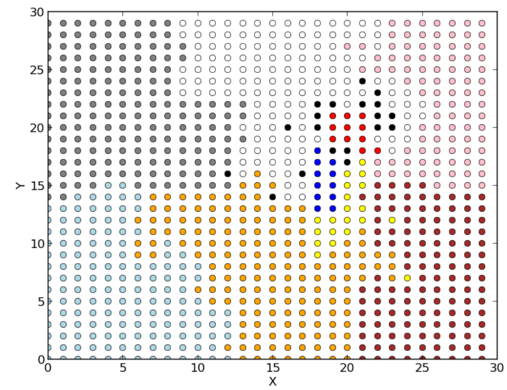

Figure 4: Division of city grid cells into clusters based on neighbourhood and violent crime intensities. Each color represents a different cluster.

1. Cluster Modelling

To predict violent crime rates, one needs to have sufficient average crime count rates. Going by the data, we have 10 years of data spread over 900 grid cells. Moreover crime intensity is very high in few particular grid cells (mentioned with red dots in Figure 4), almost 30 folds of the average crime in other areas. Hence, we needed to cluster city regions taking into account the neighbourhood factor and crime intensities. The areas prone to crime and in neighbourhood to each other are to be considered seperately treating them as a seperate cluster and balancing their above average crime count by making such clusters as small as possible. Similarly for low crime zones, the cluster sizes are larger if low or almost bo crimes occur. This was done so to provide maximum aid to the police authorities as frequent crime zones require greater vigilance and hence must be predicted with greater resolution. Henceforth, clusters were generated using the attributes of location and crime intensity and the new clustered city model is shown in Figure 4.

2. Time series prediction

Having modelled the city as a set of clusters, we need to build up the set of attributes required for the same to run

Time division	Police Distt.	Lift Coeff.
0600-0630	Tenderloin	1.6340733208
0130-0200	Central	1.5286333502
1430-1500	Tenderloin	1.383249552

Table 1: Correlation results- Time & Police Distt.

Crime	Day	Police Distt.	Lift Coeff.
Prostitution	Sunday	Park	3.404
Extortion	Wednesday	Park	3.055
Prostitution	Saturday	Park	2.510

Table 2: Correlation results- Crime, Day & Police Distt.

the prediction algorithms. The dependence of crime intensity on day would depend upon the point location of the day in the series, i.e, the *Date*, the *Day of week* at that particular day and the recent history crime trends in the region or the cluster with which the area is associated. Based upon the experimental results described in the results section, the best set turns out to be for including the past week's crime count as the parameter for capturing crime trend in the region.

Results and Analysis

Correlation among the attributes

The results of the lift measures among different attributes can be seen in Table 1 and Table 2. Table 1 shows the correlation between the attributes of time and police district. Table 2 shows the correlation between the crime category, day of the week and the police district.

Predicting crimes at a given time of a day

As explained in an earlier section for finding relation between time of day and crimes we used the modified T-month approach with an extra parameter for the time of the day. 75% percent of the data was used to train the classifier and the rest 25% (the most recent data) was used as the test set.

For the crime intensive areas we note, with considerable surprise, that crime follows a near consistent behaviour. As can be seen in the graph with varying time and month we see a neat periodic graph. This can be explained from the fact that during a day crime does not occur with equal probability. Indeed we found that morning hours like 6am have less crime rate than that at 6pm. Evening time is the most vulnerable period in the terms of number of occurences of crime incidents. Hence we get a consistent time dependent curve. However for areas with sparse crime incidents its difficult to predict the time of the day on which crime will take place and hence the results degrade.

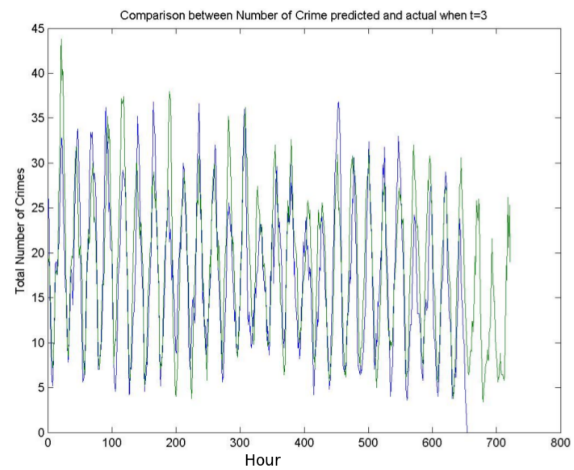

Figure 5: Comparison between total predicted crimes and actual crimes vs time of the day and months

Furthermore the results for the scenario where we wish to predict whether or not crime will occur at a particlar place without caring about the number of crime incidents are summarized in Table 3. The relative mean error denotes the difference between the actual value of number of crimes and the predicted number of crimes.

We also varied the value of k to find out the value of k that gave the best results (Figure 6). The SSE for our classifier decreased with an increase in the value of k. For the most crime intensive area, the SSE attained a minimum value at K = 42. However, it was observed that the performance did not improve much beyond $k = 10$. Hence for furthur results the value of k was set to 10.

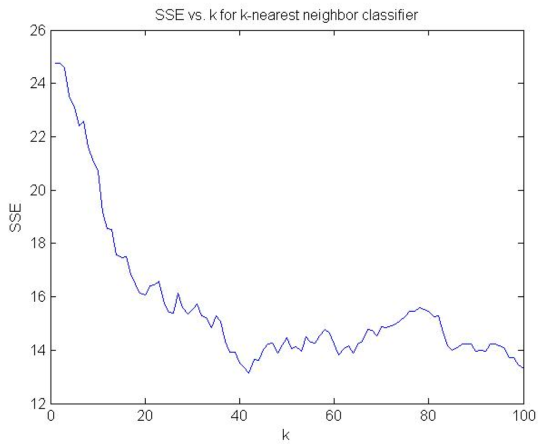

Figure 6: SSE for nearest neighbor classifiers with varying K

Prediction at a given day of the week

For predicting the possibility of crime for all days of the week, three different classifiers were tried - Naiye Bayes, Simple Logistic and Multilayer Perceptron. As described in the methodology, we used the crime data of the past T months for predicting the crime at a particular day of the week in this month. As a result, the dataset for the classifier consisted of $(117 - T)$ entries (117 being the total number of months for which the data was available) for each day of the week. The classifier's test set comprised of the 16 recent entries and the training set comprised of the remaining $(101 - T)$ entries. The **recall** and the **accuracy** of the predictions for different grid cells were calculated. Results for some grid-cells are summarized in Table 4.

We notice that recall is very low or 'not defined' in the areas having low crime rate. In these areas, the classifier mostly gives negative results (no crime), since the training data is composed of almost all negatives. Recall is not defined in the absence of crime in the testing data. For areas having high crime rate, the classifier mostly gives positive results, which is mostly true, giving very high recall values. Most of these areas witness frequent crime incidents. The prediction results for medium crime-intensive areas are also shown. The accuracy and recall obtained varies for different areas (grid cells).

3 month approach with hourly prediction

Grid-no	Recall	Accuracy	Avg Relative Error
15	0.714	0.957	0.191
618	0.923	0.802	1.986
784	0.788	0.953	1.101

3 month approach without hourly prediction

Grid-no	Recall	Accuracy	Avg Relative Error
68	0.833	0.833	2.030
157	0.763	1.0	0.111
280	0.702	0.962	0.133
637	0.884	0.821	3.546

4 month approach without hourly prediction

Grid-no	Recall	Accuracy	Avg Relative Error
16	0.892	0.961	2.468
630	0.968	1.0	6.430
637	0.961	0.806	3.589
767	0.807	0.875	2.513

Table 3: Prediction results: predicting the possibility of crime

It is interesting to note that better prediction results are obtained for Saturdays and Sundays. The results for the medium crime-intensive areas are shown in Table 5. Improvement in both recall and accuracy is seen in some areas, while for other areas no particular change is seen.

Figures 7 and 8 show the Accuracy and Recall for different classifiers at a particular grid cell for different values of T, the number of months taken as attributes. We notice that both accuracy and recall increase till $T = 6$, after which the performance remains the same/decreases. This shows that crime depends more/only on recent months and less/not on distant months.

Time series prediction

Using the methodology mentioned for time series prediction, clustering of the discrete city regions (X and Y bins) was done using hierarchical clustering with average linkage and number of clusters pre-set to 10. The attributes for clustering algorithm were X count of the grid cell, Y count and *total violent crime intensities* in the past 10 years in the grid cell. Since crime counts are very

Low crime-intensive areas

Grid-no	Recall	Accuracy
12	-	53%
155	0	67%
897	-	100%

Medium crime-intensive areas

Grid-no	Recall	Accuracy
68	0.653	40%
72	0.56	66%
163	0.581	58%
167	0.65	50%

High crime-intensive areas

Grid-no	Recall	Accuracy
624	1.0	90%
630	0.95	55%
634	0.724	51%
757	1.0	93.75%

Table 4: Prediction results - Low, medium and high crime intensive areas ($T = 5$).

Grid-no	Recall	Accuracy
68	0.67	56%
72	0.833	68.75%
163	0.581	58%
167	0.67	50%

Table 5: Prediction results for weekends - Medium crime intensive areas ($T = 5$).

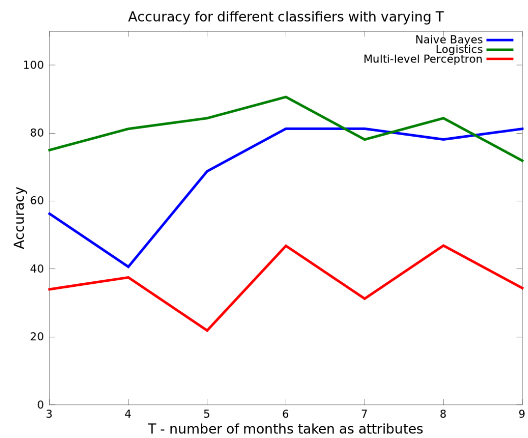

Figure 7: Accuracy for different classifiers with varying T

much unevenly distributed, repeated heirarchical clustering was used to seperate out the neighbouring crime intensive zones. There upon each cluster had different prediction algorithm operated upon it as the performances of the algorithm differ with the average crime count and the randomness in the crime occurences. Clustering the grid cells does give a better performance for prediction of crime counts in a given day as compared to non clustered as grid cells within the cluster correlate with each other on the average crime count and randomness. Also, the neighborhood factors which were considered in other approaches only considered the local neighborhood. Here, we take into account a broader area for crime prediction. The results vary with the set dependencies and are summarized in Table 6.

Further, when the prediction algorithms were performed on the various clusters, the mean absolute error and the relative absolute error were recorded. The results are ordered in the order of high crime zones to low crime zones. It was observed that in frequent number of cases, SVM based regression outperformed all major regressors including ANN(Table 7). This does establish the known fact of SVM outperforming the ANNs in time series forecasting [7]. This basically is because SVMs are trained to minimize the upper bound on generalization error while ANNs are trained for minimizing the training error. However when we see the results, ANNs perform better where crime count does not show high deviance. These are the areas where crime intensities are usually around the mean and the white noise is not too much. Test data becomes almost similar to training data and hence ANNs outperform SVMs. The problem no longer becomes identifying typical time series regression but fitting the training data.

As per the results in Table 7, the relative absolute error is quite high, especially for the low crime zones. However for the high crime zones, we do gain a comparatively better performance. This is acceptable as low crime zones have low average crime count and even slight deviances penalise the regressor to a greater extent as compared to

Attribute Set	Best Mean Absolute Error	Best Relative Absolute Error	Regressor
Date	8.2	81.4	ANN
Date, DayofWeek	7.83	79.58	ANN
Date, DayofWeek, CC(30)	7.6	77.33	ANN
Date, DayofWeek, CC(30),CC(60,30)	8.03	82.03	ANN
Date, DayofWeek, CC(7)	7.56	76.9	SVM

Table 6: This table explains how varying the dependencies affect the performance. CC(x) denotes crime count in past x days. Results shown are for cluster 4 of the clustered city model. Violent crime follows more of a weekly pattern than a monthly pattern.

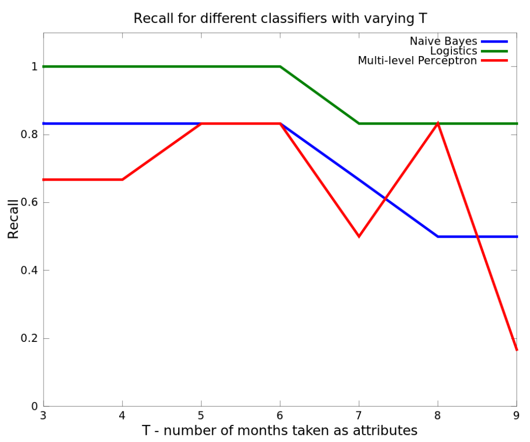

Figure 8: Recall for different classifiers with varying T

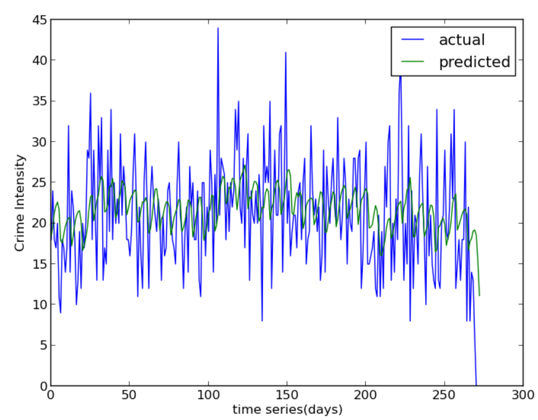

Figure 9: Time series prediction for a cluster with high average crime intensity. The changing magnitudes of white noise in prediction curve highly correlate with that of actual curve

high crime zones. But getting better predictions on high crime zones helps in the cause. The time series prediction for a high crime zone is shown in Figure 9. Though the prediction signal curve seldom meets the peaks of the actual signal curve but the changing magnitudes of white noise in prediction curve highly correlate with that of actual curve. Since such a prediction is done long before the actual time series, the noise in the prediction curves can alarm the authorities of future violent crimes in that cluster.

Conclusions

As evident from the tabular data on the lift measures, primitive inferences on the patterns related to crime category are that there are relatively frequent number of Crime incidents related to Prostitution on Weekends in the PARK Police District. Also TENDERLOIN Police District is more vulnerable to crime in the morning at around six.

Cluster Set	Best Mean Absolute Error	Best Relative Absolute Error	Regressor
9	12.08	57.33	*SVM*
3	9.47	70.77	*SVM*
2	5.27	23.66	*ANN*
4	7.53	76.58	*SVM*
7	5.73	76.67	*SVM*
5	5	66.86	*SVM*
8	5.2	67.77	*ANN*
0	4.075	62.75	*SVM*
1	5.33	96.69	*ANN*
6	2.14	94.3	*SVM*

Table 7: Results on cluster placed in the order of crime intensities. Highest crime zone first

As mentioned about crime prevention strategies from the inferences drawn from the pattern recognitions, a suitable measure to handle this pattern could be employed to prevent Prostitution incidents in the PARK police district and using law enforcements in TENDERLOIN district in the morning times.

Also, the spatial and temporal predictions provided in this report can be of great importance to law enforcements while deciding strategies. With the help of this work, a system can prevent crime by going ahead of time and predicting crime intensities in a month for a particular hour, crime intensity counts in the weekends of a month and the time series prediction for the days to come in the various regions of the city.

References

[1] H. Chen, W. Chung, J. Xu, G. Wang, Y. Qin, and M. Chau, “Crime data mining: a general framework and some examples,” *Computer*, vol. 37, pp. 50 – 56, april 2004.

[2] S. V. Nath, “Crime pattern detection using data mining,” in *LWA 2006: Lernen - Wissensentdeckung - Adaptivitt, Hildesheim, 9. - 11. Oktober 2006*, Universitt Hildesheim, 2006.

[3] C.-H. Yu, M. W. Ward, M. Morabito, and W. Ding, “Crime forecasting using data mining techniques,” *2010 IEEE International Conference on Data Mining Workshops*, vol. 0, pp. 779–786, 2011.

[4] C. Kahn, “www.npr.org/2011/11/26/142758000/at-lapd-predicting-crimes-before-they-happen/,”

[5] H. De Knegt, F. Van Langevelde, M. Coughenour, A. Skidmore, W. De Boer, I. Heitkönig, N. Knox, R. Slotow, C. Van der Waal, and H. Prins, “Spatial autocorrelation and the scaling of species-environment relationships,” *Ecology*, vol. 91, no. 8, pp. 2455–2465, 2010.

[6] L. Han, “Economic analyses of crime in england and wales.” July 2010.

[7] F. E. Tay and L. Cao, “Application of support vector machines in financial time series forecasting,” *Omega*, vol. 29, no. 4, pp. 309 – 317, 2001.