Geographical crime rate prediction

Problem statement

Its very challenging task to identify the location with high probability of crime. Although a lot of work have done to predict the type of crime. But more effort need to be put in the area of finding the geographical area or location sensitive to certain to set of crimes with the help of data mining and modeling techniques. The prediction of future crime location can be determined by Geographical data mining approaches. Appropriate analysis of spatial data can be helpful in the predicting the next location of crime.

Background

Past crime data analysis approach has been used by many researchers to predict the future crimes of similar nature. Less work could have been done in the field of predicting the next location of next crime. Crime hot spot can be detected using data modeling of history of crimes. For this there is requirement of effective modeling to process and analyze the large amount of crime data. As per literature spatio – temporal data mining can be utilized effectively to determine and predict the geographical area where the possibility of crime is more. Spare matrix analysis has been used by researchers for spatial clustering technique for the sequential crime predictions. SVM has also been used in literature to the level of crime rate. Where rate can be taken as reference parameter for determining the crime location. If given data point is data set are above some predefined threshold of crime rate then the data points may considered as member of next hotspot of crime or location. Other than these approaches, a multivariate time series clustering technique has also been used which is based on Bayesian theory.

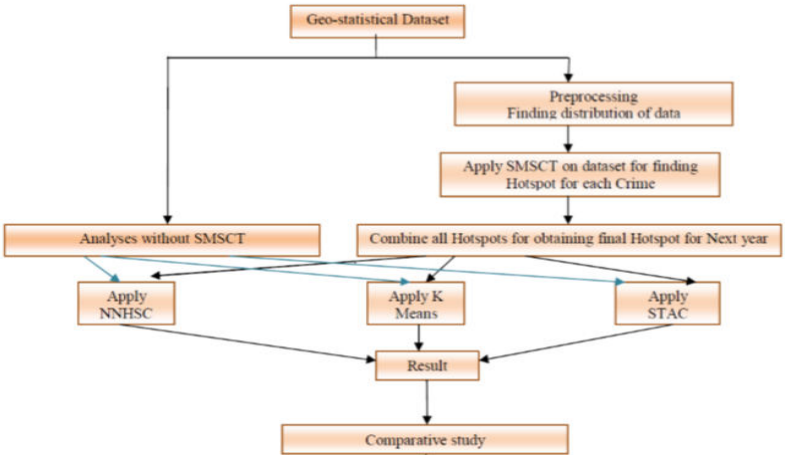
Challenges: The existing approaches identify the locations based available data of crime in terms of higher frequency of crimes i.e. density of crime. But these approaches don't consider the time and date of crime or type of crime is not taken into account. It is very challenging to predict the location of crime based on these parameters but if these parameters are also taken into account then we can even better geographical crime prediction.

Methodology

The major overall activities that need to be performed as one of approach for geographical crime prediction would be as follows:

- With use of some dataset having crime record of two different cities.
- Based on available dataset, a data mining model is prepared
- Training of data model is done with a new dataset
- Finding the relationship between existing crimes to predict the next location and apply SMCST for this purpose (Sparse Matrix analysis based spatial clustering technique).
- Apply K-means or other clustering algorithm to find the clusters.
- Crime Parameters to focus for analysis would be crime type, time during which crime happened and location of crime.

- Determine all possible patterns with some frequency and having interrelation on crime parameters.
- Then apply some classification method to predict some type of crime in specific locations using data mining approaches.



General framework for crime Prediction model

Experimental Design

Dataset considered for the proposed solution is the data set of metropolitan cities of India for minimum four years. Data is stored in form of longitude and latitude information of crime location and frequency of such incidences regarding different types of crimes. Data preprocessing is performed to extract and understand the distribution of data. For the purpose of preprocessing convex hull with standard deviation cab be deployed. Convex hull gives boundary wall around the points of distribution and deviation may describe the nature of data from mean point of view.

To find the next crime location, time series data is considered with spatial clustering base matrix analysis is applied. To predict the crime location various clustering approaches has been proposed to be used such as NNHSC (Nearest Neighbor hierarchal) , K-means and STAC(spatial analysis of crime.

To predict the crime at particular location again we need to some crime parameters as discussed earlier. The choice of location for the crime by criminals depends on many parameters. Usually similar pattern is adopted by criminals to pursue the crime. To further enhance the quality fo outcome some data reduction strategy is applied. Initial analysis gives big view of data. Statistical analysis is done on various attributes of data. Summary of data analysis can be mentioned in the following format. (Figure-1)

Attribute	Number of Distinct Values	Value
Crime_Type (nominal)	6	Assault Drug Alcohol Other crimes Public Disorder Theft White collar crime
Crime_Type_Id (numeric)	6	1: Assault 2: Drug Alcohol 3: Other crimes 4: Public Disorder 5: Theft 6: White collar crime
Crime_Month (nominal)	12	months names
Crime_Day (nominal)	7	days of the week
Crime_Time (nominal)	6	T1: 1am to 4:59 am T2: 5 am to 8:59 am T3: 9 am to 12:59 pm T4: 13 pm to 16:59 pm T5: 17 pm to 20:59 pm T6: 21 pm to 0:59 am

Figure-1

To predict the next crime location Bayesian classifier worked well to determine the next location. The result format for storing the city crime frequency pattern is shown below based on dataset.

Frequent pattern	Min-sup	Frequent pattern	Min-sup
'Capitol-hill', 'Monday', 'T5'	0.001	'Five-points', 'Thursday', 'T4'	0.001
'Capitol-hill', 'Thursday', 'T6'	0.001	'Five-points', 'Thursday', 'T5'	0.002
'Capitol-hill', 'Friday', 'T5'	0.001	'Five-points', 'Thursday', 'T6'	0.002
'Capitol-hill', 'Friday', 'T6'	0.002	'Five-points', 'Wednesday', 'T3'	0.001
'Capitol-hill', 'Saturday', 'T6'	0.002	'Five-points', 'Wednesday', 'T4'	0.002
'Capitol-hill', 'Sunday', 'T6'	0.001	'Five-points', 'Wednesday', 'T5'	0.002
'CBD', 'Monday', 'T4'	0.001	'Five-points', 'Wednesday', 'T6'	0.002
' CBD ', 'Monday', 'T5'	0.001	'Five-points', 'Saturday', 'T1'	0.001
' CBD ', 'Tuesday', 'T3'	0.001	'Five-points', 'Saturday', 'T5'	0.002
' CBD ', 'Tuesday', 'T4'	0.001	'Five-points', 'Saturday', 'T6'	0.002
' CBD ', 'Wednesday', 'T3'	0.001	Five-points', 'Sunday', T1'	0.001

Through analysis of the data for each nearby location , It was found next location which was targeted for crime have maximum wealth and population associated. From the study it is reflected dangerous nearby locations have more number of male member with age range of 21 to 29 which increase the possibility of crime occurrence.

As a conclusion we need accurate classification and clustering approaches to improve the accuracy level of location based crime prediction. Relationship of other parameters like income, age group of location can be further analyzed to have better result.