International Journal of Criminology and Sociological Theory, Vol. 9, No. 1, August 2016, 1-11.

A Geo-Statistical Approach for Crime hot spot Prediction

Sumanta Das[1]
Malini Roy Choudhury[2]

Abstract

Crime hot spot prediction is a challenging task in present time. Effective models are needed which are capable of dealing with large amount of crime dataset and prediction of future crime location. Spatio-temporal data mining are very much useful for dealing with the geographical crime data. In this paper sparse matrix analysis based spatial clustering technique for serial crime prediction model is used. Firstly, crime data are preprocessed through various distribution techniques and then sparse matrix analysis based spatial clustering technique are applied on a four years time series data from 2010 to 2014 for the major cities of India like Delhi, Mumbai, Kolkata and Chennai to find out the hotspot location for next year, after that three clustering techniques are used to grouping similar crime incident, at last cluster results obtained by original and proposed dataset are compared. The main objective of this research is applying crime prediction technique, forecast and detect the future crime location and its probability.

Introduction

Geo-statistical approach for crime analyses required past crime data to predict most probable future hot spot and time very efficiently, (Zheng et.al. 2011) and (Xue and Donald, 2003). The prediction of future crime hot spot can be done by Geographical data mining techniques (Berry et.al. 1999) and (Larose, 2005). The spatio-temporal data mining provides the platform to analyze the spatial data. It concerns with both spatial and temporal relationships two important attributes of spatial –temporal data mining are Location and Time. The area of spatio-temporal data mining is where this relationship is both defined by the spatial and temporal characteristics of the data and is extremely challenging due to the increased search space for knowledge (Roddick and Spiliopoulou, 1999).

In this paper sparse matrix analysis based spatial clustering technique for serial crime prediction model is used which is based on spatial temporal data mining and also used for predicting future crime location. Crime Hot-Spots Prediction Using Support

[1] Department of Civil Engineering, SRPEC (Gujarat Technological University), Unjha, Gujarat, India, sumanvu_27@yahoo.co.in

[2] Department of Civil Engineering, SRPEC (Gujarat Technological University), Unjha, Gujarat, India, maliniroychoudhury@gmail.com

Vector Machine for a given percentage of the data defines level of crime rate. The data points which have the crime rate above the predefined rate are members of hotspot class and data points with crime rate below the predefined rate are members of cold spot class. k-median clustering algorithm is used for this purpose. Lastly compare the result when the same percentage of the data is selected randomly (Kianmehr and Alhajj, 2006).

A Multivariate Time Series Clustering Approach for Crime Trends Prediction. In this technique, a approach for multivariate time series clustering technique based on DTW together with Parametric Minkowski model is used to consider the weightage scheme in the clustering algorithm (Chandra et al., 2008). A Novel Serial Crime Prediction Model Based on Bayesian Learning Theory. This technique introduces a novel serial crime prediction model using Bayesian learning theory. Author mainly studied the factor related to geographic report which is made by combining all geographic profiles weighted by effect functions which can be adjusted adaptively based on Bayesian learning theory (Liao et.al. 2010). Predicting the geo-temporal variations of crime and disorder (Jonathan J. Corcoran et.al. 2003). This technique introduces for crime incident prediction by concentrate on geographical areas of concern that outshine traditional policing boundaries. The result of this technique are satisfactory using artificial neural network and gamma test provide the facility to predict future crime (Shrivastav and Ekata, 2012) and (Liu et.al., 2012).

Dataset

To test the different approach used by our model, the Indian crime data set is used. The dataset has details of four types of crime in various cities of India. The location of each data point is described in the data set by Latitude and Longitudes.

Methodology

In this work, our aim is to propose the technique used for predicting the future serial crime hotspot The sparse matrix analysis based spatial clustering technique is used to find out future hotspot based on previous year time series data and after obtaining hotspot different clustering techniques are used to find out similar group of cluster. To evaluate the performance comparison of different clustering algorithm applied on original and proposed data are done (Figure 1).

- First it takes crime incident data from various location of specified land.
- Apply pre-processing techniques to find distribution and distance analysis of crime incident data.
- Apply SMSCT to find out future hotspot.
- Apply NNHSC, K-Means and STAC clustering to find out cluster.
- Compare result of clustering on original and proposed data set.

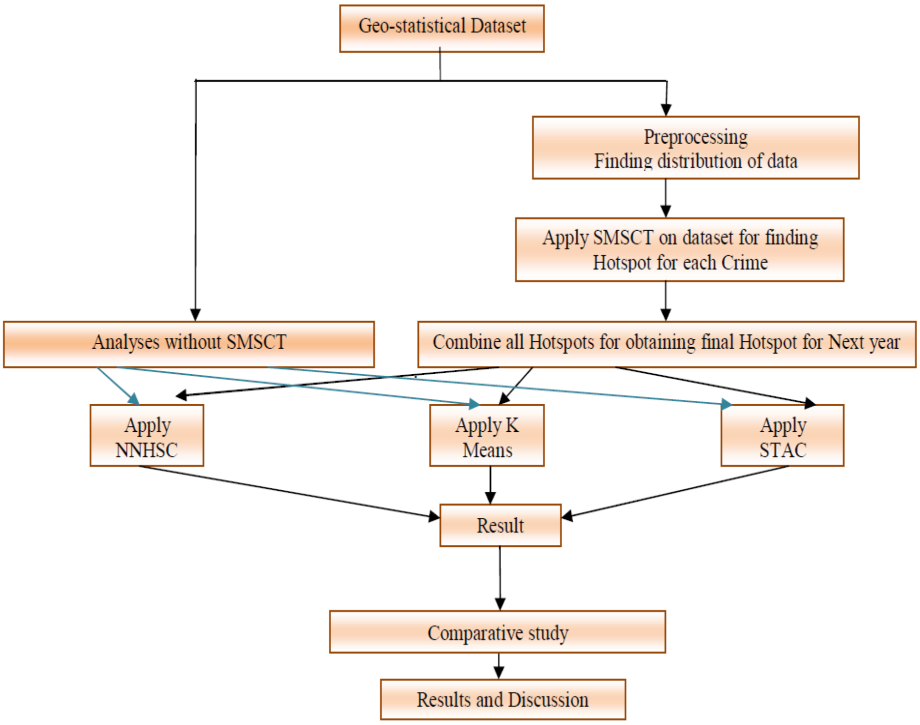
Geo-statistical Dataset
Preprocessing
Finding distribution of data
Apply SMSCT on dataset for finding Hotspot for each Crime
Analyses without SMSCT
Combine all Hotspots for obtaining final Hotspot for Next year
Apply NNHSC
Apply K Means
Apply STAC
Result
Comparative study
Results and Discussion

Figure 1: Proposed frame work for Crime Prediction model

Data Set Selection

For this work Indian crime data of four metropolitan cities from 2010 to 2014 have been collected. It consists of latitudinal and longitudinal extension and event count of four different types of crime.

Preprocessing

In this phase, data are firstly preprocessed for finding out the distribution of data. For this purpose, two preprocessing techniques i.e. standard deviation ellipse and convex hull have been used.

Standard Deviation Ellipse

The standard distance deviation shows the dispersion of the incidents around the mean center.

Convex Hull

The convex hull is a boundary drawn around the distribution of points, it represents a polygon that describes all the points in the distribution.

Sparse Matrix Analysis based Spatial Clustering Technique

To find out the hotspot location of next year based on analysis of previous years time series data spatial clustering technique based on sparse matrix analysis have been proposed.

The steps of SMSCT technique are as follow:

Step-I: Retrieve crime data from the data set in the yearly grid. (Grid is divided into cell where each cell represents the particular city).

Step-II: Add all the values of same cell from starting year grid to last year grid.

Step-III: If resultant value contains at least one event of specified crime type than it is consider as hotspot and assign the value 1 otherwise cold spot and assign the value 0 in specified cell.

Step-IV: Apply sparse matrix technique for removing cold spot from the final grid and obtaining hotspot location for specified crime type.

Step-V: Merge all the crime type hotspot location for obtaining final hotspot location of next year.

Clustering

After obtaining the final hotspot location of crime for next year, apply three type of clustering techniques for finding similar group of crime incident.

NNHSC

The nearest neighbor hierarchical spatial clustering is a constant-distance clustering that groups points together on the basis of spatial proximity. The user defines a threshold distance and the minimum number of points that are required for each cluster, and an output size for representing the clusters with ellipses.

K-Means

The K-means clustering is a procedure for partitioning all the points into K groups in which K is a number assigned by the user. This algorithm finds K seed locations in which points are assigned to the immediate cluster.

STAC

The Spatial and Temporal Analysis of Crime (STAC) is a variable-distance clustering routine. It initially associate points together on the basis of a constant search radius, but then combines clusters that overlap.

Results and Analysis

Result of SMSCT Technique

By Applying SMSCT technique on the data set based upon analysis of four years crime data(2010-2014), hotspot location are found for each type of crime which are shown in table. There are four types of value are output in each table. TRACT – Crime location, LON – Longitude, LAT – Latitude, OFFENCE – Type of offence (Table 1).

Table 1: Location based individual crime event from 2010 to 2014

Table:1				Table:2			
TRACT	**LON**	**LAT**	**OFFENCE**	**TRACT**	**LON**	**LAT**	**OFFENCE**
Mumbai	72.87766	19.07598	Murder	Mumbai	72.62471	19.07661	Robbery
Delhi	77.22496	28.63531	Murder	Delhi	77.45871	28.58532	Robbery
Kolkata	88.462	22.6263	Murder	Kolkata	88.662	22.7163	Robbery
Chennai	80.17	13.04	Murder	Chennai	80.3716	13.2145	Robbery
Table:3				**Table:4**			
TRACT	**LON**	**LAT**	**OFFENCE**	**TRACT**	**LON**	**LAT**	**OFFENCE**
Mumbai	72.77265	19.17598	Kidnap	Mumbai	72.5932	19.12361	Pickpocket
Delhi	77.32296	28.63531	Kidnap	Delhi	77.38542	28.52453	Pickpocket
Kolkata	88.512	22.5113	Kidnap	Kolkata	88.604	22.6182	Pickpocket
Chennai	80.1812	13.1147	Kidnap	Chennai	80.3067	13.1265	Pickpocket

After merging all the hotspot obtained according to different crime type, final hotspot for the next year are obtained (Table 2).

Table 2: Predicted hotspot for next year 2015

TRACT	LON	LAT
Mumbai	72.6438	19.1578
Delhi	77.3129	28.7651
Kolkata	88.5686	22.5226
Chennai	80.1573	13.0891

Result of Clustering Techniques

After obtaining the final predicted value for next year, clustering technique are applied for grouping similar type of location in a multiple group.

Result of NNHSC for Original Dataset

By Applying Nearest Neighbor Hierarchal Spatial Clustering algorithm on the original dataset ten different clusters have been found (Table 3).

Table 3: Clusters are generated by Nearest Neighbor Hierarchal Spatial Clustering algorithm on the original dataset and density is calculated accordingly

Cluster	Mean X	Mean Y	Density
1	72.78634	20.48462	0.006028
2	83.37951	24.80211	0.00478
3	86.67047	22.04364	0.005003
4	75.91047	23.30167	0.005974
5	79.76221	26.82437	0.009952
6	81.34597	21.18579	0.976893
7	71.96633	26.57735	3.24565
8	80.13478	16.16153	0.003214
9	77.62858	29.4759	1.253707
10	75.90639	17.65992	763.9437

Result of NNHSC for Proposed Data set

By Applying Nearest Neighbor Hierarchal Spatial Clustering algorithm on the proposed dataset eight different clusters have been found (Table 4).

Table 4: Clusters are generated by Nearest Neighbor Hierarchal Spatial Clustering algorithm on the proposed dataset and density is calculated accordingly

Cluster	Mean X	Mean Y	Density
1	83.39694	25.04271	0.000246
2	72.72239	20.12864	0.000216
3	79.44327	26.88842	0.000335
4	87.1877	22.8152	0.000284
5	80.22092	16.10704	0.000284
6	85.50054	20.02451	0.002817
7	81.14617	21.13436	40.754839
8	75.48415	23.3764	0.000131

Result of K-Means for Original Dataset

By Applying K-Means Clustering algorithm on the original dataset five different clusters have been found (Table 5).

Table 5: Clusters are generated by K-Means Clustering algorithm on the original dataset

Cluster	Mean X	Mean Y	Mean Sqr Error
1	79.7929	13.41402	849.990763
2	80.3733	17.00579	4142.540522
3	73.7507	18.93218	11194.63116
4	83.8357	23.41188	45204.80477
5	74.6963	24.19458	57303.87098

Result of K-Means for Proposed Data set

By Applying K-Means Clustering algorithm on the proposed dataset five different clusters have been found (Table 6).

Table 6: Clusters are generated by K-Means Clustering algorithm on the proposed dataset

Cluster	Mean X	Mean Y	Mean Sqr Error
1	83.0776	25.4357	5365.46571
2	80.4715	16.8815	3620.916487
3	82.7259	21.14759	3063.61781
4	73.1103	21.69475	4393.73457
5	79.7514	13.43788	839.497076

Result of STAC for Original Dataset

By Applying STAC Clustering algorithm on the original dataset seven different clusters have been found (Table 7).

Table 7: Clusters are generated by STAC Clustering algorithm on the original dataset

Cluster	Mean X	Mean Y	Density
1	83.5416	25.14251	0.006949
2	72.8312	21.23201	0.00488
3	76.2803	23.09007	0.008778
4	81.6846	21.47633	0.045005
5	85.5005	20.02451	0.11268
6	77.6165	27.90599	1.804891
7	79.7929	13.41402	44.00768

Result of STAC for Proposed Data Set

By Applying STAC Clustering algorithm on the data set three different clusters have been found (Table 8).

Table 8: Clusters are generated by STAC Clustering algorithm on the proposed dataset

Cluster	Mean X	Mean Y	Density
1	82.9842	24.55383	0.000153
2	74.4094	22.17456	0.000086
3	85.8538	20.37929	54.250183

Comparative Study and Discussion

Table 9: Comparison of Clusters Obtained by NNHSC, K-MEANS and STAC

DataSet	Number of Clusters		
	NNHSC	K-Means	STAC
Original Dataset	10	5	7
Proposed Dataset	8	5	3

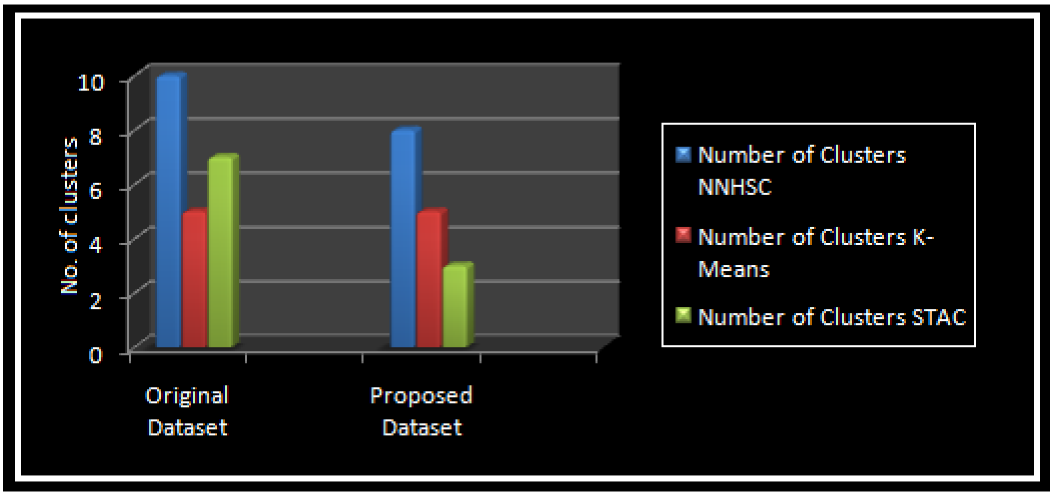

Figure 2: Graphical representation of clusters by NNHSC, K-MEANS and STAC

Table 9 & Figure 2 show the Comparative Analysis and graphical representation of Clustering Techniques Applied to Original and Preprocessed Dataset. In our experiment we obtained different results of both clustering algorithms. According to Table 9 we evaluate clusters for both original and preprocessed dataset for grouping the similar type of crime incident locations. When NNHSC applied on original dataset than number of cluster found is 10 and after processing when STAC based NNHSC applied on dataset than number of cluster found are 8. Similarly for other two clustering algorithm applied on original dataset found 5 and 7 clusters respectively and after preprocess when these two SMSCT based clustering technique is applied on dataset than number of clusters found is 5 and 3 respectively. The result found that similarity measure of similar crime incident location is increased and our results of SMSCT based clustering algorithms obtain more efficient result.

Table 10: Comparison of Clusters Obtained by SMSNAT NNHSC, K-MEANS and STAC Algorithms

Number of Clusters		
SMSNAT based NNHSC	**SMSNAT based K-Means**	**SMSNAT based STAC**
10	5	7
8	5	3

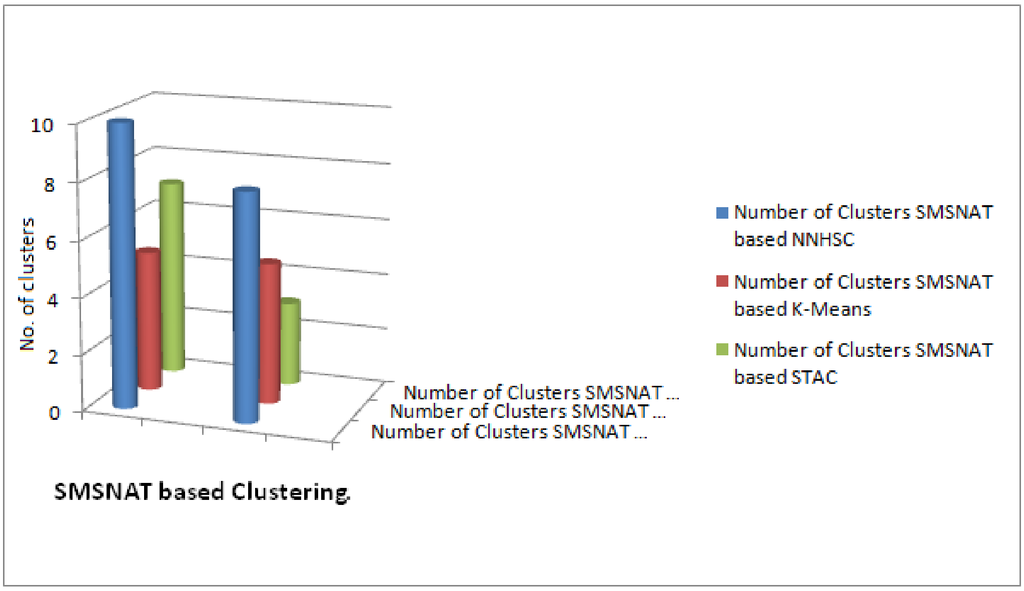

Figure 3: Graphical representation Clusters Obtained by SMSNAT NNHSC, K-MEANS and STAC Algorithms

Table 10 & Figure 3 show the Comparison of Clusters Obtained by SMSCT based NNHSC, K-Means and STAC clustering algorithm. According to the table 10 & Figure 3 when SMSCT based Nearest Neighbour Clustering algorithm is applied on preprocess dataset than number of cluster found is 8. Similarly when SMSCT based K-Means and STAC clustering algorithm is applied to preprocess dataset than number of cluster found is 5 and 3 respectively. On comparing these three SMSCT based clustering algorithm minimum number of cluster found is 3 by SMSCT based STAC Algorithm. When number of cluster is decreasing it means each cluster group more number of similar crime incident location. In our work the result found that similarity measure of similar crime incident location is increased in SMSCT based STAC clustering algorithm and STAC is best clustering algorithm for analyzing crime incident location and for predict future crime location.

Conclusion and Recommendations

In this paper a Geo-statistical Spatial Clustering Technique for Crime Prediction is represented. This technique mainly followed by geographical information of the serial crime incidents of different cities on different years. Firstly crime data are preprocessed through various distribution techniques and then sparse matrix analysis based spatial clustering technique are applied on a four year time series data and find out the hotspot location for next year, after that three clustering techniques are used to grouping the similar crime incident The approach can play an important role for analyzing and solving crime related problems and identification of hot spot in future trends.

References

Anand Kumar Shrivastav and Dr. Ekata, 2012. "Applicability of Soft computing technique for Crime Forecasting: A Preliminary Investigation" *International Journal of Computer Science & Engineering Technology*, 415-421

B. Chandra, Manish Gupta, M.P Gupta, 2008. "A Multivariate Time Series Clustering Approach for Crime Trends Prediction" *IEEE*, 892-896.

Jenn-Long Liu, Chien-Liang Chen and Hsing-Hui Yang, 2012. "Efficient Evolutionary Data Mining Algorithms Applied to the Insurance Fraud Prediction" *International Journal of Machine Learning*, 308-313.

John F. Roddick and Myra Spiliopoulou, 1999. "A Bibliography of Temporal, Spatial and Spatio-Temporal Data Mining Research", *ACM 5*, 34-38.

Jonathan J. Corcoran, Ian D. Wilson, J. Andrew Ware, 2003. "Predicting the geo-temporal variations of crime and disorder", *International Institute of Forecasters*, Elsevier, 623-634.

Keivan Kianmehr and Reda Alhajj, 2006. "Crime Hot-Spots Prediction Using Support Vector Machine", *IEEE*, 952- 960.

Larose, D. T. 2005. "Discovering Knowledge in Data: An Introduction to Data Mining", *Wiley & Sons, Inc*, 143-154.

M. Berry et al., 1999."Introduction to Data Mining and Knowledge Discovery", *Two Crows Corporation*, Third Edition, 1-39.

Renjie Liao, Xueyao Wang,Lun Li and Zengchang Qinh,2010. "A Novel Serial Crime Prediction Model Based on Bayesian Learning Theory" Ninth International Conference on Machine Learning and Cybernetics, Qingdao, IEEE, 1757-1762.

Xifan Zheng, Yang Cao, Zhiyu Ma, 2011. "A Mathematical Modeling Approach for Geographical Profiling and Crime Prediction" *IEEE*, 500-503.

Yifei Xue and Donald E. Brown, 2003. "A Decision Model for Spatial Site Selection by Criminals: A Foundation for Law Enforcement Decision Support", *IEEE*, 78-85.