

CRIME PATTERN RECOGNITION AND PREDICTION USING OPTIMISED K-MEANS AND SVM

¹M.Vijayalakshmi, ²Shivansh Bhatt, ³Himanshu Goyal^{1,2,3}SRM Institute of Science And Technology(KTR),Chennai, India¹vijayalakshmi.ma@ktr.srmuniv.ac.in, ²shivanshbhatt_bharat@srmuniv.edu.in,³himanshu_sudhirkumar@srmuniv.edu.in

Abstract: An act done by any person which is against the laws of the particular country or region is called crime. A person who does this is called a criminal. Almost without exception, people in America think that the crimes are increasing at all-time high levels. Such worries are linked to anxiety about drugs, permissive childrearing, hedonism, declining academic standards, and the collective inability to compete with the others. In Chicago the police maintain a criminal record which is accessible to the masses. The basic idea of what things are called "crimes" is that they are thought to be things that might cause a problem for another person. In this paper, we cluster the different parameters or attributes that are available with us in the dataset. After that we use SVM to do the prediction that were can the next crime happen. The output will be in the form of clusters mapped on a boxplot and the prediction will also be on a boxplot in the form of a graph. These results will help the police departments to plan their strategies according to the results obtained and prevent further crimes happening.

Keywords: Data mining, machine learning, optimized k-means, SVM, clustering, crime profiling

1. Introduction

All the government departments worldwide are giving top priority to the security and are focusing on curtailing the crime occurrence. In perpetrating the crime, the criminals have maximum utilization of the novel practices and the modernized innovation. In general, all the police departments around the world monitor their crime data and statistics and draw conclusions from that. But for humans to process all the available data into a systematic manner needs computing powers. This is where data mining and machine learning comes handy to process the large chunk of data that the departments maintain and draw conclusions based on that and take preventive measures to reduce the number of crimes.

Data mining and machine learning can be a tremendous tool for crime pattern detection, and for predictive policing in general. The police can try to stop the crimes if the crime patterns are identified and used for the planning. Without such tools, it could take

weeks or years of sifting through a database to discover a pattern, or it might be missed altogether.

In this paper, we use the optimized k-means algorithm for clustering. The observation of only a small part of data elements changing their clusters after a certain number of iterations effectively eliminates the need to re-distribute data elements. This elimination of re-distribution leads to the optimization of the running time. The data set is not in the perfect form to be used and mined for desired results, so first we preprocess the data set so that it can be mined properly through the algorithms used. Few attributes are changed from text to numeric and indexed so that the data can be used. The crime data set is inputted into the python scripts and optimized k-means is used to create various clusters. The clusters are made by taking 2 attributes at a time from the dataset. Then by using SVM, we will predict that what can be the possible location of the next crime. The results can be used by the police for making action plans and do the duties accordingly.

2. Background

In Chicago, the Chicago Police Department(CPD) [1] has the crime data set available for the public. Identifying the trends and patterns and making tactical analysis is the duty of criminal investigators across the world [2]. The safety of the communities and the prevention of crime is ensured by the analysis.

Lalitha and Suresh et.al [3] discussed about the cluster based zoning of the crime info using the WEKA tool, distributing the states in India on crime frequencies.

Cosmin Marian explored the scope of optimization of the k-means algorithm for faster runtime [4].

Several data mining techniques were discussed by Shiva Prasad [5]. The mechanism of crime solution to identify criminal patterns using k-means clustering was utilized by Nath [6].

K Means technique and decision tree algorithms were used to identify clusters and also for crime data analysis by Kadhim Swadi Al-Janabi [7].

A fuzzy clustering technique was proposed for crime analysis and also to forecast and counter the crimes by Adeyiga and Bello [8].

Semi-supervised learning approach is used for building a new software tool for data mining. It is used to help in analysis and crime investigation for various scenarios by the police departments. This particular work was done by Malathi and Santhosh Baboo [9].

An existing system is studied by Manish Gupta et. al. [10] which uses interactive queries for the criminal analysis used by the Indian Police.

Delhi Police FIR records were used for the criminal profiling by using k-means clustering. An estimation of predominant crime and information about highly sensitive areas was made by Priyanka Gera and Rajan Vohra[11].

3. Methodology

The dataset is obtained from the Chicago Police Department. The preprocessing of the data is done to decrease the runtime to generate result. The grouping of the objects into some category based on the attributes in called Clustering. We used optimized k-means algorithm for the clustering tasks and SVM for the prediction part.

The crime dataset is inputted to the python scripts as csv file. The obtained dataset has information about the ID of the criminal, whether the criminal was arrested or not, whether the crime was domestic i.e. was it committed against family member. The dataset also contains information about the beat which is the smallest regional division by the Chicago Police, Community Area, the year in which the crime occurred, latitude and longitude. All these attributes gave us immense scope for mining the dataset for useful patterns.

Decision plan make the basis for the Support Vector Machines. Different set of objects having class memberships are separated by a decision plane. A boundary separates the two decision planes. The illustration is done below where the objects are categorized into two, the Green part and the Red part. When a new object arrives into the decision plane, its parameters are checked and after that it is classified into either green or red.

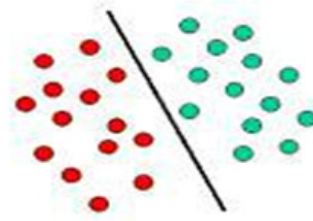


Figure 1. SVM algorithm illustration

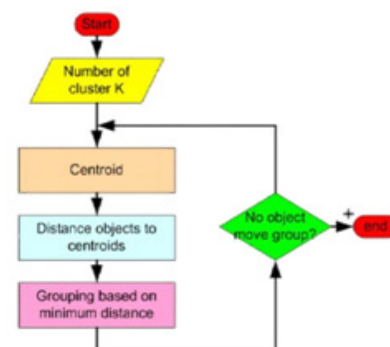


Figure 2. K-Means clustering algorithm

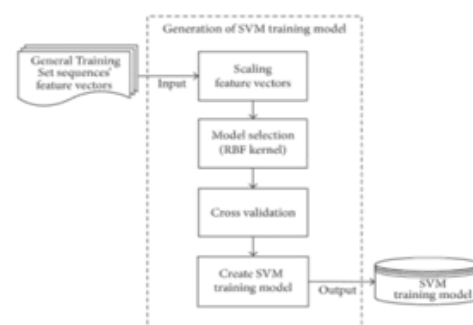


Figure 3. SVM algorithm flowchart

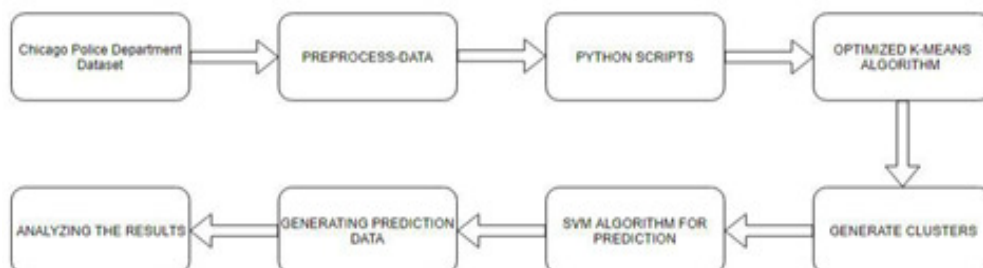


Figure 4. Methodology for the crime info

4. Experiments And Results

Table 1. Raw dataset

ID	CrimeDesc	Arrest	Domestic	Beat	District	Community	Year	Latitude	Longitude
8951354	THEFT	FALSE	FALSE	623	6	69	2012	41.75628	-87.6216
8951141	THEFT	FALSE	FALSE	1213	12	24	2012	41.89879	-87.6613
8952745	MISDEMEANOR	FALSE	FALSE	1622	16	11	2012	41.96919	-87.7677
8952223	THEFT	FALSE	FALSE	724	7	67	2012	41.76933	-87.6577
8951608	THEFT	FALSE	FALSE	211	2	35	2012	41.83757	-87.6218
8950793	THEFT	TRUE	FALSE	2521	25	19	2012	41.92856	-87.754
8950760	THEFT	FALSE	FALSE	423	4	48	2012	41.73206	-87.5648
8951611	THEFT	FALSE	FALSE	231	2	40	2012	41.79251	-87.6193
8951802	THEFT	TRUE	FALSE	1021	10	29	2012	41.86362	-87.7091
8950706	THEFT	FALSE	FALSE	1215	12	24	2012	41.89008	-87.6588
8951585	THEFT	FALSE	FALSE	1011	10	29	2012	41.86542	-87.7202
8951230	THEFT	FALSE	FALSE	734	7	67	2012	41.76506	-87.6607
8955530	THEFT	FALSE	FALSE	921	9	58	2012	41.82054	-87.6849
8950570	SHOOTING	TRUE	FALSE	825	8	66	2012	41.78455	-87.6839
8954231	THEFT	FALSE	FALSE	1834	18	8	2012	41.89349	-87.6216
8950703	THEFT	FALSE	FALSE	835	8	70	2012	41.74444	-87.6884
8954418	THEFT	FALSE	FALSE	1723	17	14	2012	41.96477	-87.7216
8951685	THEFT	FALSE	FALSE	122	1	32	2012	41.88435	-87.6311
8950586	THEFT	FALSE	FALSE	511	5	49	2012	41.71761	-87.6275

The preprocessed data about the criminal activities as obtained from Chicago Police Department is shown in Table 1. The dataset contains the information about the ID of the criminal, whether the criminal was arrested or not, whether the crime was domestic i.e. was it committed against family member. The dataset also contains information about the beat which is the smallest regional division by the Chicago Police, Community Area, the year in which the crime occurred, latitude and longitude. The crime description is also available in the dataset.

The dataset is available in MS excel format, for loading the data into python script, we save the data set

as CSV (Comma separated value) format. The Type of crime column is in alphabetical format, so all the alphabetical values were replaced with the numerical values and then indexed. In the other two alphabetical columns TRUE was replaced by 1 and False was replaced by 0. The csv data file is opened in the python script by using the csv module of python. The observation of only a small part of data elements changing their clusters after a certain number of iterations effectively eliminates the need to re-distribute data elements. This elimination of re-distribution leads to the optimization of the running time.

Table 2. Preprocessed Dataset

ID	CrimeDesc	Arrest	Domestic	Beat	District	Community	Year	Latitude	Longitude
8951354	1	0	0	623	6	69	2012	41.75628	-87.6216
8951141	1	0	0	1213	12	24	2012	41.89879	-87.6613
8952745	2	0	0	1622	16	11	2012	41.96919	-87.7677
8952223	1	0	0	724	7	67	2012	41.76933	-87.6577
8951608	1	0	0	211	2	35	2012	41.83757	-87.6218
8950793	1	1	0	2521	25	19	2012	41.92856	-87.754
8950760	1	0	0	423	4	48	2012	41.73206	-87.5648
8951611	1	0	0	231	2	40	2012	41.79251	-87.6193
8951802	1	1	0	1021	10	29	2012	41.86362	-87.7091
8950706	1	0	0	1215	12	24	2012	41.89008	-87.6588
8951585	1	0	0	1011	10	29	2012	41.86542	-87.7202
8951230	1	0	0	734	7	67	2012	41.76506	-87.6607
8955530	1	0	0	921	9	58	2012	41.82054	-87.6849
8950570	3	1	0	825	8	66	2012	41.78455	-87.6839
8954231	1	0	0	1834	18	8	2012	41.89349	-87.6216
8950703	1	0	0	835	8	70	2012	41.74444	-87.6884
8954418	1	0	0	1723	17	14	2012	41.96477	-87.7216
8951685	1	0	0	122	1	32	2012	41.88435	-87.6311
8950586	1	0	0	511	5	49	2012	41.71761	-87.6275
8951367	1	0	0	1812	18	7	2012	41.92545	-87.642
8950549	4	0	0	411	4	45	2012	41.75122	-87.5941
8950318	5	0	0	1923	19	6	2012	41.94974	-87.6543

We chose to make 5 outputs based on the clustering with the following attributes, the first output has attributes BEAT and TYPE OF CRIME which gives clustered information about the type of crime active in the particular are of Chicago. The second output is based on the attributes DOMESTIC and BEAT which gives us information about the more numbers of domestic crimes happening in the particular beat. The third output is based on the YEAR and TYPE OF CRIME, which throws highlight on past years patterns or the year in which the criminal activities were on rise. The fourth output is based on BEAT and ARREST indicating the efforts of the particular police officers of a Beat to control criminal activities for the Beat that has maximum number of arrests. The fifth output is based on the COMMUNITY AREA and TYPE OF CRIME which enables the communities in Chicago to take efforts and spread awareness about reducing the number of crimes in the particular community area. All the outputs have various clusters based on different parameters.

After the clustering is done and the output is generated, we use python's library scikit learn in the script to generate the report on the prediction.

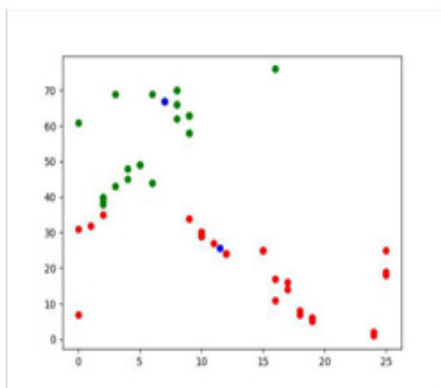


Figure 5. cluster for Beat and type of crime

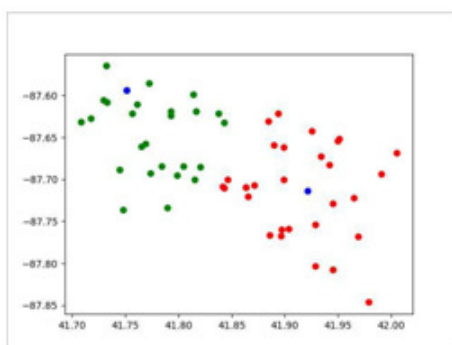


Figure 6. year and type of crime cluster

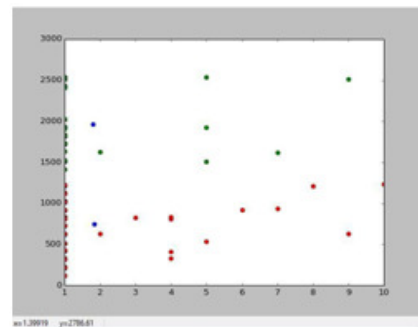


Figure 7. community area and type of crime

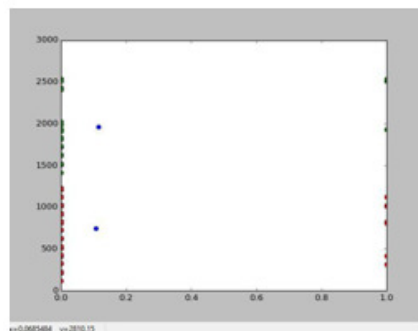


Figure 8. domestic and beat cluster

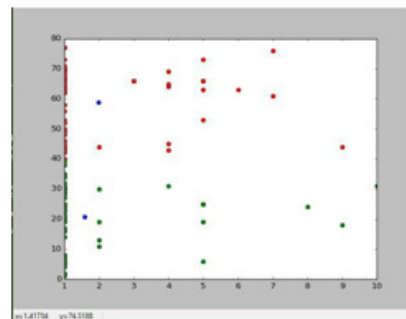


Figure 9. beat and arrest cluster

5. Conclusion And Future Work

The clustering and prediction approaches of data mining play vital role for the crime analysis which indeed helps to approach criminology in a different manner. The focus here is on using optimized k-means algorithm so that the running time can be reduced for the useful clusters to be generated and by using SVM for predicting crime pattern and drawing useful information from that.

Various output clusters are made for the different parameters so get a broader perspective on the crime

analysis that is done on the available data to generate useful information about the criminal behavior.

In future, the specific requirements can be collected from the police departments and clusters can be generated according to the needs of the department. The knowledge obtained from the generated clusters can be used to make various strategies about the crime control and optimal deployment of resources in crime prevention for future.

References

- [1] Chicago Police Department, Chicago. <http://portal.chicagopolice.org>
- [2] Lalitha Thota and Suresh Babu, Cluster based Zoning of Crime Info, IEEE, April 2017
- [3] Cosmin and Marian, An Optimized Version of the K-Means Clustering Algorithm, FedCSIS, Biohouse volume 2 issue 1
- [4] Chris Delaney, Crime Pattern Definitions for Tactical Analysis, International Association of Crime Analysts (IACA), August 2011
- [5] Shiva Prasad, Data Mining Techniques: A Short Review, Biohouse Journal of Computer Science, Biohouse volume 2 issue 1, Jan-Feb 2016
- [6] S.V. Nath, Crime pattern detection using data mining, IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, 2006
- [7] Kadhim B. Swadi al-Janabi . A Proposed
- [8] Framework for Analyzing Crime DataSet using Decision Tree and Simple K-means Mining Algorithms, Journal of Kufa for Mathematics and Computer, Vol.1, No.3, may 2011
- [9] Adeyiga J.A and Bello A.O, A Review of Different Clustering Techniques in Criminal Profiling, International Journal of Advanced Research in Computer Science and Software Engineering IJARCSSE, Volume 6, Issue 4, April 2016.
- [10] Malathi, Dr. S. Santhosh Baboo, Algorithmic Crime Prediction Model Based on the Analysis of Crime Clusters, Global Journal of Computer Science and Technology Volume 11 Issue 11 Version 1.0 July 2011
- [11] Manish Gupta, B.Chandra and M. P. Gupta, Crime Data Mining for Indian Police Information System, Computer society of India, 2007
- [12] Priyanka Gera and Rajan Vohra, City Crime Profiling Using Cluster Analysis, International Journal of Computer Science and Information Technologies IJCSIT, Vol. 5 (4), 2014
- [13] T. Padmapriya and V. Saminadan, "Distributed Load Balancing for Multiuser Multi-class Traffic in MIMO LTE-Advanced Networks", Research Journal of Applied Sciences, Engineering and Technology (RJASET) - Maxwell Scientific Organization , ISSN: 2040-7459; e-ISSN: 2040-7467, vol.12, no.8, pp:813-822, April 2016.
- [14] S.V.Manikanthan and T.Padmapriya "Recent Trends In M2m Communications In 4g Networks And Evolution Towards 5g", International Journal of Pure and Applied Mathematics, ISSN NO:1314-3395, Vol-115, Issue -8, Sep 2017.

