MANTECH Publications
Journal of Computational Intelligence Theory and Its Applications
Volume 3 Issue 1

Prediction Model for Identifying Crime Trends in Cities

Pooja Gull[1], Sparsha B M[2], Nithya A[3], Panimozhi K[4]
Students[1, 2, 3], Assistant Professor[4]
Department of Computer Science and Engineering
BMS College of Engineering, Bangalore
***Corresponding Authors' email id:** panimozhi.cse@bmsce.ac.in[4]*

Abstract

Enhancing security around our general public is one of the major challenging issues. This paper is a point by point study about the endeavours that have been made to enhance the forecast of crime intensity utilizing machine learning and big data to help the citizens to remain in a sheltered zone. In this manner, the sort of study discussed here will help reveal the crime rate of an area progressively.

Keywords: *Machine learning, Sentiment Analysis, Tweets, KDE(Kernel Density Estimation), SVM(Support Vector Machine)*

I. INTRODUCTION

Crime is a destructive behaviour that harms an individual and the whole community around him. It is an impermissible behaviour which are punishable through law. These are social acts which affects the economic status of the society. The crime issues are growing in certain cities of the world.

So, there is a obligation for a technology that can predict the occurrences of crime. Automated crime detection is much needed in the society. This will help the police departments in great extent for utilizing their resources in an efficient manner. The study and analysis on crime using computer technology, critique of the tactics targeting crime and legal strategies involved with crime, helps in predicting digital crime, digital terrorism, an information warfare.

In this paper, the usage of geographical analysis together with sentimental analysis is used on data in social networks to analyse crime spots. This helps us to detect the pattern of crimes incidents. Twitter is

Page 1-10 © MANTECH PUBLICATIONS 2018. All Rights Reserved

free social networking and micro-blogging services which provides various services to users such as sending short text updates, called “tweets".

These updates help us to infer the key information's about the surroundings when the incidents occur. Using these information's, a system is devised to extract information from the tweets of different cities that were regarded to be the most dangerous cities in the country. A geographic analysis and the sentiment analysis of data uncovered a connection among these tweets and the crimes that occurred in the corresponding cities.

Sentiment analysis techniques were implemented on these tweets to carry out the analysis and later predict the intensities of crime in particular location. These processes need to be identified and their performance can then be improved using data from previous executions.

PROBLEM STATEMENT

Automatic crime prediction tool is the major requirement for today's world for the prevention of crime. In order to achieve our goal, we set three objectives :(1) contextual analysis in twitter data to score positivity or negativity and trend in the neighbourhood using sentiment analysis (2) using Kernel Density Estimation (KDE) to analyse the distribution of crime accidents. (3) Building a predictive model using twitter-data and KDE.

RELATED WORK

I) Sentiment Analysis:

Here, the study is about the feature-based opinion summarisation of criminal tweets. The assignment is performed in two stages: Begin, with recognising the keywords of the tweets that people have tweeted. Further, each tweet is distinguished as positive or negative impression. There are technically two methodologies for finding terms in dictionary in terminology identification.

First is the symbolic approach which depends on syntactic depiction of terms, to be a specific noun phrases. Second is the factual approach which makes us the way the word making the term tend to be discovered near each other at reoccurring. The system performs the summarisation in two main steps: 'Feature Extraction' and 'Opinion orientation identification'.

Former system, given the data sources, the frame work initially downloads all the tweets, and places them in the survey database. This system , first concentrates

Page 1-10 © MANTECH PUBLICATIONS 2018. All Rights Reserved

"hot" highlights that many individuals have communicated their opinions in tweets, after which it finds the infrequent ones. Latter system, the function compresses the feelings of components into positive and negative by taking the generated features as inputs [16].

Frequent features generation; this step is done using association rule mining. This is used to find all frequent item sets. An association rule is an implication of the form A?B, where A?I, B?I, and A n B = Ø. The rule A? B holds in D with confidence c. If c% of transactions in D that support A also support B. The rule has support s in D if s% of the transactions in D contain A? B.

Mining frequent occurring phrases, each bit of data extracted above is put away in dataset called a transaction set/document. At that point the association rule miner is run, which depends on the Apriori algorithm. This algorithm works in two stages, initial step is to discover all frequent item sets from an arrangement of exchanges that satisfy a user-specified minimum support.

Feature Pruning also known as 'Compactness pruning', method checks features that contain at least two words, called feature phrases, and removes the words that are likely to be meaningless. Redundancy Pruning mainly focuses on removing redundant features that contain single words. 'Opinion Sentence Orientation Identification', consists of two steps: (1) Using bootstrapping technique and WordNet, each word in the opinion word list identifies its semantic orientation. (2)Based on the dominant orientation of the opinion words in the sentence it decides the opinion orientation of each sentence.

II) Prediction:

Prediction is carried out by using sentimental polarity and historical crime record as variables. Our goal is to predict future crime on certain areas of the city. For predicting the future crime, we are interested in the trend of sentiment within each neighbourhood along with sentiment polarity.

The data available from twitter has a time stamp attached to it. This kind of data is known as 'time-series data'. The current ways to deal with this issue, thrown in two gatherings: Feature Filters and Feature wrappers. For the feature selection phase, the former modelling tool can be used.

Page 1-10 © MANTECH PUBLICATIONS 2018. All Rights Reserved

To select the final set of features they basically try to use some statistical properties (ex. Correlation). The latter approach includes the modelling tool in the selection process. It is an iterative process where each set of features is tried with modelling tool and its results are recorded.

The decision of models was predominantly guided by the way that these techniques are well known by their ability to handle highly non-linear regression problems. Detailed approaches to this domain would necessarily require a large comparison of more alternatives. Two such alternatives are listed be: Artificial Neural Network (ANNs) and Support-Vector machine (SVM).

(1) As ANNs have ability to deal with highly nonlinear problems they are used in financial forecasting and are formed by a set of computing units (the neurons) linked to each other. (2) SVM is the modelling tools that can be applied to both regression and classification tasks. SVMs have been seeing expanded consideration from various research groups in perspective of their fruitful application to a few areas and furthermore their solid hypothetical foundation [21].

The fundamental idea behind Support Vector Machine is that of mapping the original data into another, high-dimensional space, where it is conceivable to apply straight models to get an separating hyper plane, for instance, isolating the classes of the problem, in the case of classification tasks.

DATA COLLECTION

The primary data sources for implementing the described abstract are twitter data and Chicago crime incident data that shows the historical trends of all the crime incidents occurred in Chicago. Twitter is one of the social media filled with billions of status updates around the world.

Tweets contain lots of attributes like city, country, time, profile, status and many more. Lots of conclusions can be made through this data such as earthquake detection, mental and health informatics by studying populations health, crime detection and so on. These analyses can be carried by knowing the geo location of the users.

But the researches show that the number of the geo-tagged location are very less and can be said as up to 1% of the total tweets. It is challenging to collect the tweets from

Page 1-10 © MANTECH PUBLICATIONS 2018. All Rights Reserved

a particular location for the accurate analysis. Here, the data is being collected from the locations using the hash tags by using the twitter API, tweets are extracted as JSON objects that includes the text along with the metadata such as location, time, user identification number, user profiling. Another data source is taken from the Chicago data portal website. The Chicago police department updates by keeping the track of the crime incidents committed. They also include latitude, longitude, and area, type of crime, description and time of the crime incident.

Modeling

Key feature to settle crime problems is by studying where the crime takes place. This forms the base for recognizing crime traits through geographical location.

Beginning with the hotspot mapping, Kernel Density Estimation is regarded as the best approach for this. This is the analytical process which helps us in identifying the hotspot areas, basic form of crime prediction. There exists flexibility in choosing the bandwidth and grid cell size which are the parameters to this function.

It produces smooth surface map showing variations in the study areas. The data collected from online portal is spatially analysed for determining crime and non-crime points. The following results are obtained after the KDE analysis; it produces crime density along with latitude and longitude (shown in figure 1).

The figure 2 shows graphical representation of the output produced through kernel density estimation.

1	1.323577e-10
2	1.264901e-10
3	1.283488e-10
4	1.385847e-10
5	1.574588e-10
6	1.850331e-10
7	2.134721e-10
8	2.367660e-10
9	2.496755e-10
10	2.475076e-10

Figure 1: Output of KDE (Input-Chicago data portal website)

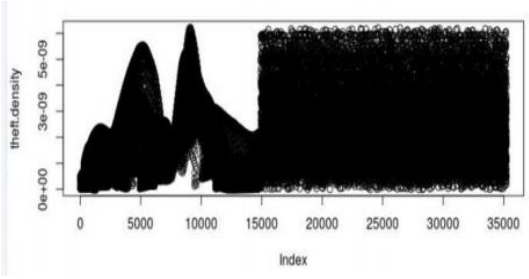


Figure 2: Graphical representation of KDE

Now, as KDE helps us in getting the crime hotspot areas, this can be used for collecting tweets from the crime hotspots and its neighbouring cities. The computation equation for KDE is as shown

Page 1-10 © MANTECH PUBLICATIONS 2018. All Rights Reserved

in eq1: .eq1 we calculate crime density at each point p.

$$x(1,t)(p) = k(p,h) = \frac{1}{Ph}\sum_{j=1}^{P} K(\frac{\|p-p(j)\|}{h})$$

We set the following parameters:

'h' as bandwidth parameter
'P' as total number of crime incidents that occurred
'j' indicates single crime point from the data of crime period
'K' is standard normal density function
||.|| is Euclidean norm
'p(j)' actual crime point.

And thus derive the distribution of crime incidents. Initiating with the collection of geo-tagged tweets from hotspot areas within the Chicago boundary, shape file of Chicago boundary is read to form the grid the map.

Then, the collected tweets are merged with the shape file and later on sentiment analysis is being carried out.

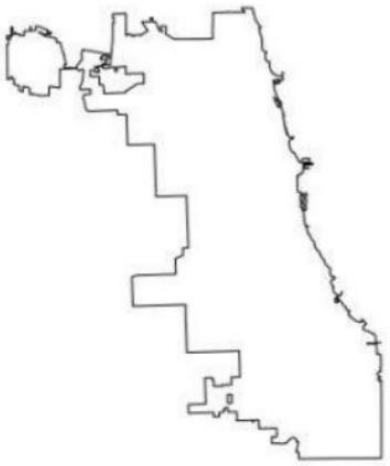


Figure 3: Chicago shape file

Further, sentiment analysis is being carried out by using 'sentiment lexicon dictionary' and 'polarity'. The positive and negative words are analysed for the sentiment analysis and classified accordingly. Polarity is being calculated in the following fashion.

Firstly, each word in the document was assigned the score using 'lexicon dictionary'. The words with negative essence are assigned negative polarity and the words with the positive essence are assigned with positive polarity. In this manner, polarity is being calculated.

The emoticons, as they have important information about the emotions of the user, they are replaced by the English words by using plyrDictionary package in R. Beside sentiment polarity, interest also lies in the

Page 1-10 © MANTECH PUBLICATIONS 2018. All Rights Reserved

trends in the trends of neighborhood. Trend index is being calculated as the change in percentage from one period to another in the given series.

This is accomplished by the inbuilt delta function in R. The formula for the trend index is shown in eq 2.

$$V_i = \left\{ \frac{P_{i-j} - P_i}{P_i} \right\}_{j=1}^{k}$$

The trend index is created by previewing k periods trend of polarity. Below are the graphs of trend index of each neighbourhood. The graphs for trend index are shown in figure 4.

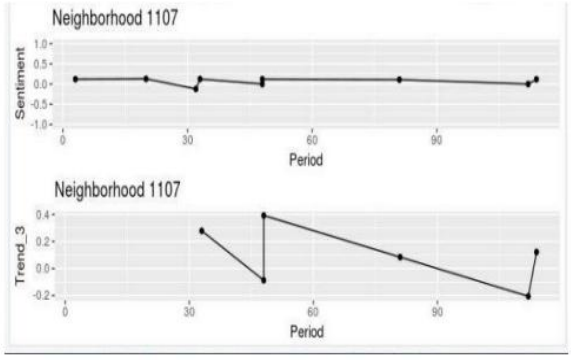


Figure 4: Polarity score and 3-day trend of neighbourhood 1107

For Prediction, SVM (Support Vector Machine) is used. Support Vector Machine is one of the best techniques to be used as it provides more accurate results as proved before.

The goal is to use KDE (Kernel Density Estimation) and the sentiment features to predict the future theft incidents. The process is carried out by setting a binary random variable to predict whether crime will occur or no (1 is taken as crime point and 0 as non-crime point). The coefficients are determined during training the model. For prediction, the crime density is calculated using KDE function.

Then tweets are processed to calculate the polarity score and trend index. SVM uses estimated density, twitter features to predict the incidents for the future period. There are two different kernels that can be used for better prediction, one is linear and other is radial kernel. Both of them give different set of outputs, among which radial is found to be more effective.

Tunning an algorithm or a machine learning technique can be thought of a process in which the parameters that impact the model is optimized to enable the algorithm to perform better. Each model is tunned once, for first model the error range is set to seq(0,1,0.1) with cost ranging from 2^2 to 2^9. In second tunning, the error range is set to seq (0,

Page 1-10 © MANTECH PUBLICATIONS 2018. All Rights Reserved

0.2, and 0.01) with same cost as previous tunning model.

It can be observed that radial kernel gives better performance than linear kernel when the error limit is set to fewer values. The tuned result for radial kernel is found to be 0.6117627. So, radial SVM can be said to be better predictor function. Figure 6 shows the output of prediction. Figure 6: Output of prediction

CONCLUSIONS

In this paper, we conclude that a violation/crime patterns can be identified, by monitoring the online social media.

Crime is meant to happen anywhere and at any-time. Previous designed statistics were unable to predict the strength/intensity of crime in a particular location accurately. But precise results can be taken when it involves soc ial networking data's. The outcome from geographic data analysis gathered through various tweets gives a vivid pattern on the criminal trends that had happened in several cities.

An enhanced sentiment analysis algorithm would help in bifurcating a sinister murderer from tweets within a confined locality. Our survey and proposed methods focuses on the model's ability to improve the forecast performance with the standard hot-spot (KDE) model on theft incidents. These results indicate not only potential benefits for the efficient allocation of policing resources but also a rewarding line for future research.

ACKNOWLEDGEMENT

We acknowledge our sponsors TEQIP-III [TECHNICAL EDUCATION QUALITY IMPROVEMENT PROGRAMME] of the MHRD, Government of India through BMSCE, Bangalore

REFERENCES

I. April 2017, Volume 4, Issue 04 JETIR (ISSN-2349-5162) JETIR1704063 Journal of Emerging Technologies and Innovative Research (JETIR) www.jetir.org 257 ENHANCE ALGORITHAM TO PREDICT A CRIME USING DATA MINING 1HARDI. M. PATEL, 2RIPAL PATEL 1 P.G.Scholar, 2Asst. Professor, 1Computer Engineering Department, 1 Silver oak college of engineering and technology, Ahmedabad, India

II. http://ieeexplore.ieee.org/document/8101496/(Distantly Supervised Lifelong Learning for Large-Scale

Page 1-10 © MANTECH PUBLICATIONS 2018. All Rights Reserved

Social Media Sentiment Analysis, Year: 2017, IEEE)

III. http://ieeexplore.ieee.org/document/8058315/ (A survey of sentiment analysis techniques, Year: 2017)

IV. http://ieeexplore.ieee.org/document/8093492/ (Sentiment analysis on Twitter data with semi-supervised Doc2Vec, Year: 2017)

V. http://ieeexplore.ieee.org/document/8117884/ (Multi-lingual sentiment analysis of Twitter data by using classification algorithms, Year:2017)

VI. http://ieeexplore.ieee.org/document/7951748/ (A survey on sentiment analysis methods and approach, Year: 2017)

VII. http://ieeexplore.ieee.org/document/8057508/ (Social Influence Algorithms and Emotion Classification for Prediction of Human Behavior: A Survey , Year:2017)

VIII. http://www.sciencedirect.com/science/article/pii/S2090447914000550 (Sentiment analysis algorithms and applications: A survey, Year:2014)

IX. http://ieeexplore.ieee.org/document/6921628/ (CRIMETRACER: Activity space based crime location prediction, Year: 2014 , IEEE)

X. https://scholarsmine.mst.edu/cgi/viewcontent.cgi?article=8320&context=masters_theses(This is 17th one) Year: 2014

XI. Toward Predictive Crime Analysis via Social Media, Big Data, and GIS Anthony J. Corso, Claremont Graduate University Gondy Leroy, University of Arizona, Tucson Abdulkareem Alsusdais, Claremont Graduate University. University of Arizona, Tucson Abdulkareem Alsusdais, Claremont Graduate University.

XII. Twitter Rest API Rate Limiting in v1.1. In Twitter Developers, The Free Service for Developers to extract data from Twitter. [Online].Available :https://dev.twitter.com/docs/rate-limiting/1.1

Page 1-10 © MANTECH PUBLICATIONS 2018. All Rights Reserved

XIII. Wang, X., Brown, D.E.: The spatio-temporal generalized additive model for criminal incidents. ISI (2011)

XIV. Automatic Crime Prediction using Events Extracted from Twitter Posts Xiaofeng Wang, Matthew S. Gerber, and Donald E. Brown Department of Systems and Information Engineering, University of Virginia {xw4u,msg8u,brown}@virginia.edu

XV. Chainey, S and S. Uhlig, L. Tomposn. 2011. "The utility of hotspot mapping for predicting spatial patterns of crime." Security Journal 21, pp. 4 – 28.

XVI. Hu, M and Liu, B. 2004 "Mining opinion features in customer reviews." National Conference on Artificial Intelligence.

XVII. "Polarity Score (Sentiment Analysis)." 2015. http://www.inside-r.org/packages/cran/qdap/docs/polarity Accessed:March 20, 2015.

XVIII. http://ieeexplore.ieee.org/document/7982289

XIX. https://www.ideals.illinois.edu/bitstream/handle/2142/73457/229_ready.pdf?sequence=2 (Toward Predictive Crime Analysis via Social Media, Big Data, and GIS

XX. Torgo, Luis. 2010. "Data Mining with R: Learning with Case Studies" Boca Raton: Chapman and Hall, pp. 107 – 109.

XXI. G. King and L.Zeng, "Logistic Regression in Rare Events Data", University of Harward, Cambridge, MA, 2001, pp.137-163.

Page 1-10 © MANTECH PUBLICATIONS 2018. All Rights Reserved