Available online: March 24, 2018

Commun.Fac.Sci.Univ.Ank.Series A2-A3
Volume 60, Number 1, Pages 11-20(2018)
DOI: 10.1501/commua1-2_0000000107
ISSN 1303-6009 E-ISSN 2618-6462
http://communications.science.ankara.edu.tr/index.php?series=A2A3

CRIME PREDICTION USING SOCIAL SENTIMENT AND SOCIO-FACTOR

SAKIRIN TAM AND Ö. ÖZGÜR TANRIÖVER

Abstract. Crime prediction becomes very important trend and a key technique in crime analysis to identify the optimal patrol strategy for police department. Many researchers have found number of techniques and solutions to analyze crime, using data mining techniques. These studies can help to speed up and computerize the process of crime analysis processes. However, the pattern of crime is flexible, it always changes and grows. With social media, user posts and discusses event publicly. These textual data of every user has contextual information of user's daily activities. These posts generate unstructured data that can be used for data prediction. As shown by previous research, twitter sentiment enable to predict crime in Chicago, United States. However, existed model on crime prediction was incorporating the use of socio factors. Therefore, the study aims to model crime prediction using social media content with additional socio-factors. The research approach is consisted of a combination of sentiment analysis from Twitter and social-factors with Kernel Density Estimation. Lexicon-base methods will be applied for sentiment analysis, and the model evaluation is measured with the help of logistic regression.

1. Introduction

Over the last decade, the exponential of social media development allows users to publish and access information and idea freely around the globe. With ease and effective use of social media, the number of Twitter users is grown from 140 million user with 340 million tweets or messages in 2012 [1] to 313 million users and 1 billion tweets or actively post on twitter [2]. Recent researches show that Twitter can be used for decision support as an ideal data source; users publicly discuss and distribute topics, emotions and events in real time with precise and valuable information called hashtag. Hashtag contains hidden information that can detect events and trend of topic [3]. As result, many researches used Twitter as predictive analytic data source to predict large-scale events in natural disaster [4], election [5] and crime [1] [6].

Received by the editors: November 11, 2017; Accepted: January 20, 2018.
Key word and phrases: Crime prediction, Data mining, Social media, Sentiment analysis, Socio-Factor.

2018 Ankara University

Crime detection and prediction become very important trend and a key technique in crime analysis to identify the optimal patrol strategy for police department. Many researches have found number of techniques and solutions to analyze crime, using data mining techniques. These studies can help to speed up and computerize the process of crime analysis processes. However the pattern of crime is not static, it always change and growth [7].

With social media, users post and discuss event publicly. These textual data of every user's post has contextual information of user's daily activities. These posts generate unstructured data that can be used for crime prediction. Previous researches had shown that twitter sentiment enable to predict crime in Chicago, United States. However, existed models on crime prediction present only sentiment analysis on twitter content with Kernel Density Estimation of historical crime. The models are lack of applying socio factors. In contrast, crime prediction model will produce more significant if socio-factor is added to sentiment analysis and KDEs.

Therefore, the study aims to fulfill the gap mention above by considering socio-factor especially gender, age and education level to be significantly influent factor that could lead a person commit crime. The paper is structured as follows: Section 2 discusses related work. Selection 3 discuses the development of prediction model. Section 4 presents model evaluation, and section 5 is the conclusion.

2. Related study

2.1. Crime Prediction

Hot-spot maps are a traditional method of analyzing and visualizing the distribution of crimes across space and time [8]. Relevant techniques include Kernel Density Estimation (KDE), which fits a two-dimensional spatial probability density function to a historical crime record. This approach allows the analyst to rapidly visualize areas with historically high crime concentrations. Future crimes often occur in the vicinity of past crimes; therefore making hot-spot maps is a valuable crime prediction tool. KDEs are a model that has been applied for crime detection base on history of crime data. It displays higher risk in area for higher density of crime. KDEs consist of ArcGIS, MapInfo, R, and CrimeStat III that is applicable for crime hotspot detection. To improve the performance of crime detection and prediction, KDEs recently use spatial, temporal, and social media data [9][1][10].

More advanced techniques like self-exciting point process models also capture the spatiotemporal clustering of criminal events [11]. These techniques are useful but carry specific limitations. First, they are locally descriptive, meaning that a hot-spot model for one geographic area cannot be used to characterize a different geographic area. Second, they require historical crime data for the area of interest, meaning they cannot be constructed for areas that lack such data. Third, they do not consider the rich social media landscape of an area when analyzing crime patterns.

Researchers have addressed the first two limitations of hot-spot maps by projecting the criminal point process into a feature space that describes each point in terms of its proximity to, for example, local roadways and police headquarters [12]. This space is then modeled using simple techniques such as generalized additive models or logistic regression. The benefits of this approach are clear: it can simultaneously consider a wide variety of historical and spatial variables when making predictions; furthermore, predictions can be made for geographic areas that lack historical crime records, so long as the areas are associated with the requisite spatial information (e.g., locations of roadways and police headquarters). The third limitation of traditional hot-spot maps; the lack of consideration for social media has been our objective in our research.

2.2. Prediction via social media using sentiment analysis

In a forthcoming survey of social-media-based predictive modeling, Kalampokis et al. identify seven application areas represented by 52 published articles [13]. As shown, researchers have attempted to use social media to predict or detect disease outbreaks [14], election results [15], macroeconomic processes (including crime) [16], box office performance of movies [17], natural phenomena such as earthquakes [18], product sales [19], and financial markets [20]. These researches primarily use the technique of sentiment analysis. Researchers employ semantic analysis on the contextual contents of each tweet and draw the predictive response of the selected group of people.

A primary difference between nearly all of these studies and the present research concerns spatial resolution. Whereas processes like disease outbreaks and election results can be addressed at a spatial resolution that covers an entire city with a single prediction, criminal processes can vary dramatically between individual city blocks. The work by Wang et al. comes closest to the present research by using tweets drawn from local news agencies [16]. The authors found preliminary evidence that such tweets can be used to predict hit-and-run vehicular accidents and breaking-and entering crimes; however, their study did not address several key aspects of social-media-based crime prediction. First, they used tweets solely from hand-selected news

agencies. These tweets, being written by professional journalists, were relatively easy to process using current text analysis techniques; however, this was done at the expense of ignoring hundreds of thousands of potentially important messages. Second, the tweets used by Wang et al. were not associated with GPS location information, which is often attached to Twitter messages and indicates the user s location when posting the message. Thus, the authors were unable to explore deeper issues concerning the geographic origin of Twitter messages and the correlation between message origin and criminal processes. Third, the authors only investigated two of the many crime types tracked by police organizations, and they did not compare their models with traditional hot-spot maps.

In our study, we addressed these limitations. We applied sentiment analysis to evaluate the polarity of tweets. We also integrated socio-economic factors into the model with sentiment polarity and historical crime record as explanatory variables. Taking full advantage of all these features, we are able to develop more accurate prediction on future crime incidents.

3. Prediction model

To construct crime model, Frist, we define training set of crime density, data from twitter and data of socio-factors. These data contain location where crime is occurred (latitude and longitude) within a specific training window (for example: 1January 2018 – 31 January 2018). The location of spatial points are designed whether they have crime or not within 200qm. Therefore, a city of training set is divided into multitude of smaller region with length of 200qm. All crime points are then assigned to small sectors (neighborhoods). At this step we have a binary classifiers to determine crime and non-crime point. Second, we collect data from twitter with a specific latitude/longitude and training window as mention above. The data contain two parts; first tweet post and second socio factor-that we can derive from user profile such as gender, age and employment status. Tweet will be computed to find polarity score of sentiment analysis. Last, we combine socio-factor, tweet polarity score with crime density. Let us have variable as following:

Response Variable
$X_t(p)$ = Crime or non-crime point (0 is non-crime and 1 is crime point)

Explanatory variable
Explanatory variable
$f_1(p)$ = Density of past crime
$f_2(p)$ = Tweet polarity score (-1 to 1)

$f_3(p)$ = Trend of polarity score
$f_4(p)$ = Gender
$f_5(p)$ = Age
$f_6(p)$ = Employment status

3.1. Historical crime density $f_1(p)$

To calculate historical density of a particular type of crime (for example theft crime) at p point, let take $f_1(p)$ be Kernel Density Estimation (KDE) at point p:

$$f_1(p) = k(p,h) = \frac{1}{Ph}\sum_{j=1}^{k} K\left(\frac{\|p-p_j\|}{h}\right) \quad (1)$$

From Eq.(1), we have p is the point that we need to calculate the crime density, set h to be parameter (or bandwidth) to control the smoothness of density, P is total number of a crime (theft crime from the example) that occur in the city, set j to index a single point of crime, K is the function of density (we can use standard normal density function), $||.||$ is the Euclidean norm, p_j is the actual point of crime (theft) in the city. To help calculation of $k(p,h)$, we can use ks packaged in R software, and we can also use Hpi function to get values of h.

3.2. Information messages from Twitter $f_2(p)$

Messages from Twitter will be collected within a specific latitude/longitude and training window of time. At this point, the spatial sectors (neighborhood) lay down to 1000 x 1000 meter. Within defined spatial sectors (neighborhood) above, we are able to find polarity score of sentiment base on given tweet message from users in each sector. Polarity function and Sentiment Lexicon Dictionary will be used to perform for this purpose.

Once we find sentiment polarity, we then find sentiment trend in each neighborhood, because it intuitively cause higher risk of crime. To measure sentiment trend, let take $f_3(p)$ as trend of polarity score, X_i is the percentage of change between previews k of period's polarity and today's polarity is:

$$V_i = \sum_{j=1}^{k}\left(\frac{p_{i-j}-p_j}{p_j}\right) \quad (2)$$

If the period is changed over 10%, we get sum of the change for $f_3(p)$ as follows:

$$=T_i = \sum_v \{v \in V_i : v > 10\% \cup v < 10\%\} \tag{3}$$

If the value of T is positive, the polarity of preview period is greater than polarity of current period. And if T is negative, the polarity of current period is greater than preview period.

3.3. Logistic Regression

In order to predict future crime that may occur, we define $X_{t+1}(p)$ to determine whether crime will occur in p neighborhood on the next day. Full form of logistic regression is shown in Eq.(4).

$$log\left\{\frac{Pr[x_{t+1}(p)=1]}{1-Pr[x_{t+1}(p)=1]}\right\} = \beta_0 + \beta_1 f_{t,1}(p) + \cdots + \beta_6 f_{t,6}(p) \tag{4}$$

$(\beta_0 \ldots \beta_6)$ can be calculated by using maximum likelihood function of historical crime (for example theft crime).

Our objective is to develop a model to predict crime incident using Twitter sentiment polarity, socio-factor on KDE model. We first estimate historical crime density on each neighborhood p at time t using KDE as mention in Eq.(1). We then calculate polarity of tweets and its trend at each neighborhood p as mention in Eq.(2) and Eq.(3). Logistic regression model is then performing density estimation, feature from Twitter and socio-factor to forecast the probability of crime incident on next day t+1 in each point.

4. Model evaluation

To measure the performance of logistic regression, we use surveillance plots. Surveillance plots measures a crime as percentage of capture on y-axis of prediction window per percentage captured area for a particular crime prediction. If the value of sentiment polarity is increasing, the higher rate a crime could be occurred. The function of capture crime and captured area is written as:

$$\%\ \text{Area Surveilled} = x\% = \frac{\sum_{i=1}^{s} A_i}{\sum_{i=1}^{n} A_i} \tag{5}$$

$$\%\ \text{Crime Captured} = y\% = \frac{\sum_{i=1}^{s} C_i}{\sum_{i=1}^{n} C_i} \tag{6}$$

where $\{A_n\}$ is the capture area sorted by point p, $\{c_i\}$ is the capture crime with captured area, and s is the number of capture area. From this equation, we derive function of surveillance curve (AUC) as:

$$\text{AUC} = \int_{-\infty}^{\infty} y(A) * x'(A) dA \tag{7}$$

Meaning that, in order to evaluate prediction model we use AUC of prediction performance then compare with benchmarked logistic regression model.

5. Conclusion

In this paper we have presented the model development of crime prediction using sentimental contents, socio-factor with crime density estimation. Few researches have explained the correlation between crime and sentiment and none has taken into account these together in model construction. Therefore, this research fulfills this gab. However, this research present only initial model development, we are planning to further improve it with real training data set.

References

[1] Matthew, S. G., Predicting crime using twitter and kernel density estimation, *Decision Support Systems,* 61 (2014), 115–125.

[2] Twitter, TWITTER USAGE / COMPANY FACTS, Retrieved from http://www.twitter.com Retried on November 1, 2017

[3] Salim A. and Omer, E., Cybercrime Profiling: Text mining techniques to detect and predict criminal activities in microblog posts, *International Conference on Intelligent Systems: Theories and Applications* (SITA), (2015) 1-5.

[4] Vieweg, S., Hughes, A. L., Starbird K. and Palen, L., Microblogging during two natural hazards events: what twitter may contribute to situational awareness, *SIGCHI Conference on Human Factors in Computing Systems*, (2010) 1079–1088.

[5] Tumasjan, A., Sprenger, T. O., Sandner, P. Q. and Welpe, I. M., Predicting elections with twitter: What 140 characters reveal about political sentiment. *ICWSM,* 10 (2010),178–185.

[6] Xiaofeng, W., Matthew S. G. and Donald, E. B., Automatic crime prediction using events extracted from twitter posts. Social Computing, *Behavioral-Cultural Modeling and Prediction*, (2015) 231–238.

[7] Sathyadevan, S., Devan M. and Surya, S., Gangadharan, Crime analysis and prediction using data mining, *Networks Soft Computing (ICNSC),* (2014) 406–412.

[8] Chainey, S., Tompson, L. and Uhlig, S., The utility of hotspot mapping for predicting spatial patterns of crime, *Security Journal*, 21(2008), 4–28.

[9] Caplan, J.M. and Kennedy, L.W., Risk terrain modeling compendium. *Rutgers Center on Public Security*, Newark, (2011).

[10] Mohammad A.B. and Matthew, S.G., Predicting Crime with Routine Activity Patterns Inferred from Social Media. *International Conference on Systems, Man, and Cybernetics – SMC*, (2016), 1233-1238.

[11] Mohler, G.O., Short, M.B., Brantingham, P.J., Schoenberg F.P. and Tita, G.E., Self-Exciting Point Process Modeling of Crime. *Journal of the American Statistical Association,* 106 (2011), 100-108.

[12] Xue, Y. and Brown, D.E., Spatial analysis with preference specification of latent decision makers for criminal event prediction, *Decision Support Systems*, 41 (2006), 560–573.

[13] Kalampokis, E., Tambouris, E., and Tarabanis, K., Understanding the predictive power of social media, *Internet Research*, 23 (2013).

[14] Culotta, A. and Huberman. B., Towards detecting influenza epidemics by analyzing Twitter messages, *Proceedings of the First Workshop on Social Media Analytics, ACM*, (2010) 115–122.

[15] Franch, F., Wisdom of the crowds 2: 2010 UK election prediction with social media, *Journal of Information Technology & Politics,* 10 (2013), 57–71.

[16] Wang, X., Brown, D. and Gerber, M., Spatio-temporalmodeling of criminal incidents using geographic, demographic, and Twitter-derived information, *Intelligence and Security Informatics.* Lecture Notes in Computer Science, IEEE Press, (2012).

[17] Asur, S. and Huberman, B., Predicting the future with social media, *IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, IEEE*, (2010) 492–499.

[18] Earle, P.S., Bowden, D.C. and Guy, M, Twitter earthquake detection: earthquake monitoring in a social world, *Annals of Geophysics*, 54 (2012).

[19] Choi, H,. and Varian, H., Predicting the present with Google Trends, *The Economic Record*, 88 (2012), 2–9.

[20] Bollen, J., Mao H., Zeng, X., Twitter mood predicts the stock market, *Journal of Computational Science,* 2(2011), 1–8.

Current Address: Sakirin TAM: Department of Computer Engineering, Ankara University, Ankara 06830, TURKEY
E-mail Address: *kirin.it@gmail.com*
ORCID: https://orcid.org/0000-0003-4103-1797

Current Address: Ö. Özgür TANRIÖVER: Department of Computer Engineering, Ankara University, Ankara 06830, TURKEY
E-mail Address: *tanriover@ankara.edu.tr, tanriover*@yahoo.com
ORCID: https://orcid.org/0000-0003-0833-3494