Machine Learning Based Threat Prediction Algorithm for Citywide Real time Heat map Generation

Prof. Varsha Dange[1], Amol Pawar[2], Sagar Bhandge[3], Siddharth Prajapati[4], Riya Daryanani[5], Vaishnavi Javalkar[6]

[1] *Professor, Vishwakarma Institute of Technology*

[2,3,4,5,6] *Student, Vishwakarma Institute of Technology*

***Abstract*- The premise is to collect the numerical subjective data produced by individual end users, the device's location, as well as universal data such as Time of the year, time of the day, and apply a machine learning predictive algorithm to generate a heatmap overlapping the city or any other geographical entity simulating a predictive threat/safety level of a given area or route. Based on the data provided to the algorithm, the algorithm can suggest statistically safest routes for travelling as well as visualize the route onto the map**

INTRODUCTION

Crimes against women are a commonplace in India as well as the world. As a growing economy, a growing portion of the workforce is being occupied by female workers. With the increase in the number of women taking up the workforce and the amount of time they spend outside, there is a net increase in the number of exploitable e scenarios for perpetrators. With no preventative measures deployed to counter the vulnerability of the situation, crimes against women are likely to increase in proportion. The premise of the project is to develop a machine learning algorithm to create a pattern recognising system. This system will use the algorithm to intake relevant data produced by the end users. The algorithm will find patterns and co-relations within the data. This will enable the system to produce statistically reliable predictions for threat simulations and route suggestions. As the system works in real time, it has an instantaneous utility. The system has two components.

a. Android Application

b. Application Server

The android application installed onto the end user smartphones will perform two major tasks.

1. Read user generated rating and device generated location and transfer it to the application server.
2. Read the heat map produced by the server and display it to the user.

The Server hosts the Machine learning algorithm as well as the database. The processing of the data by the algorithm is performed by the server-side processor, meaning that the hardware bottleneck of the end user's device's hardware performance.

DATASET

A. Rating

The end user will produce a subjective rating in the form of a score regarding how safe they are feeling in that instant. The score will be out a number from scale of 1 to 10. This numerical data set is the basis of the database. Being a subjective dataset, which is obtained from the end user, there is a possibility of incorrect data points entering the system. Data cleaning methods such as outlier detection and treatment

B. Location

The application installed onto the smartphone will read the precise location and send the data to the application server.

C. Time

The server will associate the rating and the location data with a specific time of the day derived by the location of the device and the serve-side system time. The advantage of using the server-side system time is that it is not possible for the profile to have wrong data regarding the time, since the end user may change the time on their local device but the server-side data is completely reliable.

Day and month

Like time, the server will also associate the specific day of the week and month of the year with each rating generated by the user.

METHODOLOGY

The possibility of a crime or an attack is dynamic ever changing variable that is co-*-related to numerous environmental variables. As such, it is possible to make a calculated, educated estimate of the possibility of an incident at a given place, at a given time of the day, on a given day of the week, and during a given month of the year. The machine learning algorithm attempts therefore harvest the data regarding all of the above-mentioned variables that can affect the odds of an incident. i.e. If users at a given specific location rate lower at a certain time of the day, as compared to another time of the day, the algorithm can associate the time and location with a high threat level. Or if a given route is rated higher during summer than winter, perhaps due to the increased traffic during summer, the algorithm will take the month of the year into consideration when calculating the safety rating of the route in real time. In the above given example, the month of the year is an environmental variable. Such numerous variables are taken into account when calculating a safety rating for each route and area at any given time. Based on this safety rating of each route and area of the city. The safety rating then is used to create a heat map which is overlaid on the Google map. The Google map API is incorporated onto the application for this purpose.

FUNCTION

The function of the system can best be explained through a scenario. Consider the following scenario. In the Pune city, User A wants to go from Katraj chowk to Pimpri-Chinchwad.at 2 in the morning. The user has two possible routes.

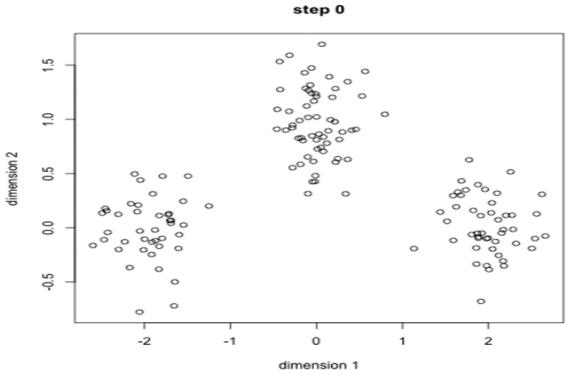
Route1: Through Swargate and the greater city Center.

Route2: Through Sinhgad road and national expressway

The algorithm reads the starting point and destination. The algorithm then binds this data with the current time, day of the week, and month of the year. Based on this a data profile is created in real time. The profile is then placed in the data cluster of the historic data generated throughout the existence of the system since deployment. The data cluster contains all the profiles. Each profile is constructed in a same format as the other. As such, the profiles created in real time i.e. one for each route, are placed in the data cluster. Based on their place in the data cluster, their safety a rating is derived. This process takes into account the different areas through which the route with take the user, as well as all the other environmental and broader variables available to the algorithm. Based on the safety rating the safest route is recommended to the end user by overlaying the route onto the map. The can also visualize the relative threat level of the entire city divided into granular areas, where each grain is a profile produced based on the data provided by the end user throughout the entire operation of the algorithm. This allows the end user to interact with a visual and graphic interface that is easy to understand, intuitive, and does not expect the end user to familiarize themselves to a new data format, as the data is in the overlaid onto a Google maps API, an interface that android users are well familiar with. This has the added benefit of seamless transition for navigation without any noticeable difference from the perspective of the end user. This makes the end user to much more likely to migrate from their usual navigation application to the one created to also keep their safety in mind.

IMPLEMENTATION

The system uses the android device as a mere interface without making it perform the bulk of the processing. The end user device is a mere tool for data gathering and displaying the end results i.e. recommended routes for the travel. The data gathering, profile building, data clustering and the pattern recognition, as well as the safety rating calculation is all done the by Server. This prevents the system from being bottlenecked by the hardware limitation of the end user device.

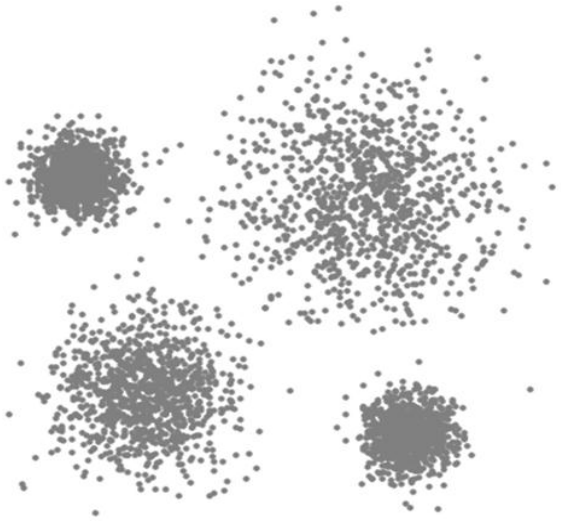


K means clustering [7]

Each data point is classified by computing the distance between that point and each group center, and then classifying the point to be in the group whose center is closest to it.

Based on these classified points, we recompute the group center by taking the mean of all the vectors in the group.[7]

K means clustering is an algorithm that enables the system to sort, classify, and divide the profiles by their similarity to each other.



Mean shift clustering[7]

In contrast to K-means clustering, there is no need to select the number of clusters as mean-shift automatically discovers this. That's a massive advantage. The fact that the cluster centers converge towards the points of maximum density is also quite desirable as it is quite intuitive to understand and fits well in a naturally data-driven sense. The drawback is that the selection of the window size/radius "r" can be non-trivial.[7]

AKNOWLEDGEMEMNT

This project is supported by Vishwakarma Institute of Technology. We thank our faculty from Vishwakarma Institute of Technology who provided insight and expertise that greatly assisted the research, although they may not agree with all of the interpretations/conclusions of this paper. We thank Prof. Varsha Dange for assistance with the algorithm implementation as well as regular guidance. We are also immensely grateful to the esteemed faculty members of the institution for their comments on an earlier version of the manuscript, although any errors are our own and should not tarnish the reputations of these esteemed persons.

REFERENCES

[1] Reference Model for Learning Analytics M.A. Chatti, A.L. Dyckhoff, U. Schroeder, and H. Thüs Informatik 9 (Learning Technologies), RWTH Aachen University {chatti, dyckhoff, Schroeder, thues}

[2] ACM SIGMETRICS Performance Evaluation Review April 2014 https://doi.org/10.1145 /2627534.2627557

[3] Learning to Predict Rare Events in Event Sequences Gary M. Weiss* and Haym Hirsh Department of Computer Science Rutgers University New Brunswick, NJ 08903

[4] Prediction System of Burning through Point (BTP) Based on Adaptive Pattern Clustering and Feature Map Wu-shan Cheng Computing Center, Department of Intelligent Robotics, Shanghai University of Engineering Science, Shanghai, China.

[5] Regressing Heatmaps for Multiple Landmark Localization Using CNNsChristian PayerEmail authorDarko ŠternHorst BischofMartin Urschler Institute for Computer Graphics and VisionGraz University of Technology GrazAustria

[6] Exploiting machine learning techniques for location recognition and prediction with smartphone logsDepartment of Computer Science, Yonsei University, 50 Yonsei-ro, Seodaemun-gu, Seoul 120-749, South Koreahttps://doi.org/10.1016/j.neucom.2015.02. 079

[7] George Seif, https://towardsdatascience.com/the-5-clustering-algorithms-data-scientists-need-to-know-a36d136ef68