Validando domicilios mediante técnicas de clustering

Daniel Negrotto and Federico Bertero

Findo, 25 de Mayo 306 Piso 2, CP 1002 CABA, Buenos Aires, Argentina
{dnegrotto, fbertero}@findo.com.ar

Abstract. Los grandes volúmenes de información y, el tiempo de respuesta esperado por los usuarios que consumen algún tipo de plataforma basada en la toma de decisiones a partir de datos, hacen de la validación automática de estos un proceso clave. Es decir, hoy en día se espera que un sistema sea capaz de verificar información brindada por un usuario, consultando diversas fuentes y dando una respuesta en un rango de tiempo relativamente corto.
En este trabajo, se busca abordar este desafío computacional cuando la información declarada por el usuario se corresponde con domicilios en los que desarrolla alguna actividad y, la fuente a utilizar para la verificación son los datos a los que se tiene acceso de su dispositivo móvil.

Keywords: Clustering · Geoinformation · Home-validation.

1 Introducción

Los servicios que utilizan localizaciones ofrecen valiosas aplicaciones para usuarios con dispositivos móviles. Para acceder a estos servicios, los usuarios aceptan el compartir su ubicación y permiten así, el uso de estos datos donde se considere necesario. Los registros de ubicación, cuando se analizan, pueden revelar tanto patrones de comportamiento de los usuarios [1, 2], como lugares mas visitados por los mismos [3].

Nuestro interés reside en utilizar los registros de ubicación que reportan los usuarios a través de sus dispositivos móviles con la finalidad de verificar datos relacionados a domicilios declarados por el mismo.

En este articulo se describe el proceso de elección y aplicación de diversos tipos y métodos de clustering, buscando el realizar esta verificación mediante un procedimiento flexible, y que, siendo capaz de dar una respuesta en un periodo corto de tiempo, sea también robusto. Se necesita flexibilidad ya que la solución debe ser capaz de adaptarse a la naturaleza heterogénea asociada a las densidades de los datos a utilizar. A su vez, se debe minimizar las consecuencias de tratar con ruido e información redundante -como suele suceder con registros de ubicación tomados de dispositivos móviles-, y con la necesidad de dar una respuesta en un periodo corto de tiempo -haciendo que la consulta a diversas fuentes para fortalecer la verificación sea acotada-.

2 Trabajos relacionados

En [1], Hu et al. estudian cómo inferir la localización del hogar de una persona a partir del contenido y de diversos atributos que se pueden extraer de sus publicaciones en la red social Twitter.

En [2], Krumm colecta los registros de ubicación producidos por equipos GPS instalados en los automóviles de un conjunto de 172 personas durante dos semanas. Con estos, y mediante cuatro algoritmos heurísticos, describe diversas formas de identificar el domicilio de vivienda de cada una de las personas del conjunto. En [3], Furletti et al., utilizando el mismo método que Krumm para obtener datos, buscan inferir qué tipo de actividad realizaron cada una de estas persona durante sus días. Para esto, se identifican puntos de interés o destinos más visitados dentro de las trayectorias recorridas, y se machea a estos con categorías de actividades. En [4], Liao et al. buscan también estimar categorías asociadas a las actividades realizadas por un grupo de personas, esta vez usando equipos de GPS portátiles para la toma de datos y redes de Markov para la inferencia.

En [5], Starner utiliza un dispositivo portátil GPS para tomar registros de ubicación de una persona durante cuatro meses y un modelo predictivo basado en cadenas de Markov para calcular, dado un punto inicial y estimando otros puntos significativos, las probabilidades de próximos trayectos a realizar por el individuo.

3 Metodología

Por cada usuario a analizar se dispone de una serie de coordenadas -latitud y longitud- asociadas a domicilios declarados por el usuario y un conjunto de registros de ubicación tomados por el dispositivo móvil del usuario cada hora. Cada uno de estos registros de ubicación está compuesto por una latitud, una longitud y una marca de tiempo que describe el momento en que fue tomado. Nuestro desafío es, para aquellos usuarios sobre los cuales disponemos de una cantidad significativa de estos registros de ubicación, comprobar si los domicilios declarados son válidos.

Cada validación se realiza de manera individual. Se comienza tomando un usuario y agrupando sus registros de ubicación bajo criterios de cercanía que discutiremos a continuación. La idea de estos grupos es identificar las actividades más significativas de un determinado usuario teniendo en cuenta la cantidad de tiempo que le dedica y sin tener una noción precisa de la actividad en sí. Finalmente, consideramos válido el domicilio declarado si la georreferencia asociada a este pertenece a uno de los grupos de ubicaciones definidos para el usuario.

Para la partición del conjunto de registros de ubicación de un usuario en grupos (clusters) se comenzó considerando que solo existen dos categorías de actividades lo suficientemente importantes: la que se lleva a cabo en el domicilio de vivienda y la que se lleva a cabo en el domicilio laboral. Es decir, en esta primera instancia, todo registro de ubicación debía pertenecer al cluster asociado

al domicilio de vivienda o al cluster asociado al domicilio laboral. Esta suposición se basa en que estos son los lugares donde una persona promedio pasa la mayor cantidad de tiempo del día. Dado que el número de clusters a formar es fijo, se optó por el algoritmo *K-Means* [6] para el particionado, utilizando como función objetivo a minimizar la suma de la distancia cuadrática entre cada registro y el centroide de su cluster.

Este primer enfoque muestra algunas debilidades. En primer lugar, el que sea necesario fijar el número de clusters genera un problema al momento de generalizar el clustering buscando identificar un número no conocido de actividades habituales de una persona. Luego, el utilizar un algoritmo de particionado no nos permite identificar puntos que se podrían considerar ruido, es decir, puntos que no se corresponden con actividades habituales de la persona. También se debe decir que *K-Means* no es un algoritmo ideal para trabajar con registro de ubicaciones ya que minimiza la varianza y no la distancia geodésica -lo que genera una distorsión en latitudes lejanas al Ecuador- y tampoco es el indicado en casos donde los clusters puede no tener una distribución convexa.

Para lidiar las desventajas de este primer enfoque se utilizó el algoritmo de clustering basado en densidad *DBSCAN* [7]. Este algoritmo nos permite construir clusters, basándose en dos parámetros: la máxima distancia que puede existir entre dos registros de ubicación para ser considerados de un mismo cluster (ϵ) y la mínima cantidad de registros que debe tener un cluster ($minPts$). *DBSCAN* nos permite identificar regiones que satisfagan la mínima densidad separadas por regiones de menor densidad y puntos considerados ruido. De esta manera, tenemos un enfoque mucho más flexible para identificar las actividades más significativas de un determinado usuario: no es necesario fijar el número de estas de antemano, sino que mediante los parámetros ϵ y $minPts$ se nos permite indicar la forma que consideramos deberían tener los clusters que las representan.

La desventaja que posee el uso de *DBSCAN* es que requiere que todos los clusters significativos presenten densidades similares. En el dominio de nuestro problema las densidades de los clusters varían según la persona a analizar, y dentro de los registros de ubicación asociados a una persona determinada también se pueden encontrar distintas distribuciones de densidad.

El último enfoque abordado se basa en el uso del algoritmo *OPTICS* [8]. Este algoritmo se puede considerar una generalización de *DBSCAN* donde se relaja el parámetro ϵ de un único valor a un rango de valores. A partir de la distancia entre registros de ubicaciones *OPTICS* crea un diagrama de alcanzabilidad que, a continuación, es utilizado para separar los clusters de densidades diferentes respecto del ruido. De esta manera, ofrece mayor flexibilidad en el afinamiento de los clusters detectados.

Si bien este último enfoque nos permite obtener clusters con diversas densidades, y con una configuración inicial más simple al no tener que lidiar con el parámetro ϵ, *OPTICS* no discrimina de forma directa los datos de entrada en clusters. La salida de este algoritmo es un diagrama de distancia de accesibilidad, y los clusters a definir quedan sujetos a la interpretación que se le da a este diagrama.

4 Resultados computacionales

Para evaluar la complejidad y tiempo de ejecución de cada uno de los enfoques presentados, se agruparon 132 usuarios elegidos al azar en categorías según la cantidad de registros disponibles por cada uno de ellos (entre 10 y 3000). De aquí se obtiene que los tiempos de validación de un usuario no superan el segundo cuando los métodos *K-Means* y *DBSCAN* son utilizados. En el caso de *OPTICS*, los tiempos de ejecución son mayores pero solo varían entre uno y dos minutos cuando el algoritmo es aplicado sobre usuarios con entre 200 y 3000 registros de ubicación asociados. Con respecto a cantidad de validaciones obtenidas por cada uno de los enfoques presentados, se obtuvo que: utilizando el primer enfoque, 71 de los 132 domicilios fue validado (aprox. 53%); utilizando el segundo enfoque, 116 de los 132 domicilios fue validado (aprox. 87%); y finalmente, utilizando el tercer enfoque, 99 de los 132 domicilios fue validado (aprox. 75%).

5 Conclusiones y trabajo a futuro

Como se dijo en la sección anterior, el segundo enfoque utilizado (el uso de *DBSCAN* como algoritmo de clustering) es el que mejores métricas presenta. Sin embargo, por lo estudiado y por la naturaleza de los datos utilizados, se considera que trabajando de forma precisa en la interpretación de los diagramas de alcanzabilidad generados por *OPTICS*, se podría lograr una mejora considerable en el tercer enfoque. Dado esto y que la cantidad de datos con los que se busca formar clusters no es suficientemente grande, se espera en un futuro cercano el poder incorporar este último enfoque al proceso de validación actual.

References

1. T. Hu and J. Luo and H. Kautz and A. Sadilek: Home Location Inference from Sparse and Noisy Data: Models and Applications. 2015 IEEE International Conference on Data Mining Workshop (ICDMW), pp. 1382-1387 (2015)
2. Krumm, J.: Inference Attacks on Location Tracks. Pervasive Computing **6**, 127–143 (2017)
3. Furletti, B. and Cintia, P. and Renso, C. and Spinsanti, L.: Inferring human activities from GPS tracks. Proceedings of the 2nd ACM SIGKDD International Workshop on Urban Computing, pp. 1–8 (2013).
4. Liao, L. and Fox, D. and Kautz, H.: Location-Based Activity Recognition using Relational Markov Networks. IJCAI International Joint Conference on Artificial Intelligence, pp. 773–778 (2015).
5. Starner, T.: Using GPS to learn significant locations and predict movement across multiple users. Personal and Ubiquitous Computing 7(5), pp. 275–286 (2003)
6. Lloyd, Stuart P: Least squares quantization in PCM. Information Theory, IEEE Transactions 28(2), pp. 129–137 (1982)
7. Ester M, Kriegel HP, Sander J, Xu X: A Density-based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, pp. 226–231 (1996)
8. Ankerst M, Breunig MM, Kriegel HP, Sander J: OPTICS: Ordering Points to Identify the Clustering Structure. SIGMOD. 1999 28(2), pp. 49–60 (1999)