Advances in Computer and Communications, 2022, 3(2), 57-69
https://www.hillpublisher.com/journals/acc/
ISSN Online: 2767-2875

Wireless Network Analysis and Optimization Based on the Social Media Data

Haijing Zhang

Department of Electronic and Electrical Engineering, The University of Sheffield, UK.

How to cite this paper: Haijing Zhang. (2022) Wireless Network Analysis and Optimization Based on the Social Media Data. *Advances in Computer and Communication*, 3(2), 57-69.
DOI: 10.26855/acc.2022.12.002

Received: October 28, 2022
Accepted: November 25, 2022
Published: December 30, 2022

***Corresponding author:** Haijing Zhang, Department of Electronic and Electrical Engineering, The University of Sheffield, UK.

Abstract

With the development of 5G, the number of mobile communication users is growing rapidly. The requirements for network communication quality are becoming increasingly high. Then the demand for wireless network optimization, especially based on social media data, rises at a super-linear rate. This paper aims to improve the performance of wireless network quality. To do this, we first collect social media data (Twitter). Clustering algorithms are one kind of machine learning (ML). We then apply two established clustering methods: K-means and Density-Based Spatial Clustering of Applications with Noise (DBSCAN) to achieve data classifier. Furthermore, we discuss the clustering results of these two algorithms and the deployment of base stations using it. Finally, the results show that it is helpful for operators to design the coverage of wireless network.

Keywords

Wireless networks, data analytics, clustering, K-Means, DBSCAN

1. Introduction

1.1 Background & Literature review

The number of internet websites and online services continues to rise, ushering in the "big data" era. However, the rise of such big data sets creates significant issues. Most research on big data analysis has been carried out. In 2014, Nawsher Khan et al. [1] published a paper in which they described traditional data analysis methods and some challenges in these approaches due to data complexity.

The big data era has also driven the development of wireless networks. Surveys such as that conducted by Suzhi Bi et al. [2] have demonstrated a scalable wireless systems for big data traffic in 2015. In introduction to big data analytical tools in wireless networks, they identified how to extract big data characteristics to design wireless networks. Nevertheless, there are issues with this wireless network system that have not been further investigated. For example, it is critical to create effective complexity reduction techniques and develop protected data processing measures in order to keep local data confidential.

Social media data is also a form of big data. One study by Michel Ballings et al. [3] reported an expert system in 2016. It is an effective tool for increasing network size on Facebook, but this study did not focus on causal relationships. In another major study, Daniel G. Costa et al. [4] found that an integrated approach named wireless sensor networks using tweets to detect and classify critical events, in order to optimize wireless sensor networks.

Recently, machine learning is one of the most rapidly developing technology areas and is at the key of artificial intelligence and data science, contributing to the development of a wide range of industries [5]. In 2018, a recent study by Weijie Qi et al. [6] involved a big-data-driven sentiment framework, which is to map consumer satisfaction towards wireless services using geo-tagged twitter data. They applied three machine learning methods to build a sentiment classifier. It has conclusively been shown that proposed SSW considerably enhances NLP performance.

In a study investigating machine learning, Yuanwei Liu et al. [7] have published an article about machine learning framework in 2019. In their review of big data analysis and machine learning, some applications in wireless networks was also presented. In this article, they applied big data analysis according to machine learning to predict the requirements of mobile users. To date, a number of studies have highlighted machine learning that are associated with wireless networks.

In view of all that has been mentioned so far, one of the most well-known tools for analyzing data is machine learning. It is also a very significant approach to develop wireless networks based on social media data.

1.2 Aims and objectives

Clustering is a form of ML on a group of samples with unknown category labels, using some algorithm to divide them into categories. K-means and DBSCAN are clustering algorithms, which are effective methods for data classification. The main purpose of this paper is to apply K-means and DBSCAN to analyze Twitter data in Barcelona, to achieve the clustering of the mobile network hotspots. Besides, it can help operators develop the deployment of cells and then optimize wireless network quality.

1.3 Overview of the paper structure

This paper is organized as follows: Following the introduction, section 2 briefly presents the impact of COVID-19. Section 3 illustrates methodology. In this section, two machine learning approaches: K-means and DBSCAN are introduced in detail. This part also shows analysis tool and data. Section 4 focuses on the results and evaluation of this paper. It shows the distribution of hotpots in Barcelona and analyze the deployment of wireless network. In section 5 states the conclusion. Finally, the paper ends with future work in section 6.

2. Impact of Covid-19

The lockdown caused by Covid-19 has influence on data collection. It means there is no data from local operators and then we purchase data from Twitter. This process affects the following steps. Then, we cannot communicate face-to-face with instructors and then keep contact with instructors by online meeting. Compared with the PID document, the objective of this research is changed to use different machine learning methods to analyze hotpots. We determine to programming in MATLAB language at the beginning but there are some problems with hardware. This obstacle could be solved by using university computer. However, we just do research remotely due to Covid-19. Therefore, we change to use python to complete this study.

3. Methodology

In this section, we will briefly describe the reason for applying social media data, the main classification of machine learning and two clustering algorithms.

3.1 Analysis Tool

In this paper, python is utilized as tool. Python is an object-oriented, interpreted, general-purpose, open source scripting language. Initially, it is easy to learn because there are few keywords and a simple structure. Second, python code is more clearly defined. Then, the major advantage of python is that it has an extensive library and is very compatible with Windows [8]. These are reasons why python is applied in this research.

3.2 Social Media Data

The enormous popularity of social media has accelerated the development of many types of data. There is a greater emphasis on the need for real-time data analytics because of the increased data collection [9]. Twitter is one of social media and it has grown in popularity in recent years. In Twitter, people share plenty of real-time data containing information about everyday events. The amount of Twitter has reached remarkable proportions and is expected to rise more in the coming years [7]. By 2018, the overall number of monthly active Twitter users had reached 330 million, and the total number of tweets sent per day had surpassed 500 million [6]. Twitter has gotten a lot of interest from the data analysis field.

Twitter data has strong correlation with actual wireless traffic demand [10]. It has shown that analysis of twitter is beneficial for the deployment of wireless network. Therefore, Twitter data is used to analyze in this paper.

In this paper, Twitter geo-location information is purchased from Twitter company.

3.3 Machine Learning

Machine learning is one of artificial intelligence tools based on the foundation that machines can learn from experience. It is a technique of data analysis. Machine learning has been effectively implemented in numerous fields. Yaohua Sun et al. [11] mentions an extensive range such as resource management and networking.

Compared to traditional methods, machine learning has the following advantages. It is able to enhance the performance of network by learning useful features from input data. Then, algorithms for networking based on machine learning can adapt well to the dynamic environment. Furthermore, each node in the network has the ability to optimize its transmission power using machine learning. Finally, Yaohua Sun et al. [11] uses examples of machine learning as evidence that it enables wireless systems solving quickly a new problem.

In a study investigating machine learning, Yuanwei Liu et al. [7] reported that the useful information can be extracted from tweets using machine learning. As an example, users are unhappy about its wireless services when they visit a famous tourist site. In order to address this, we can improve the average tourist experience by temporarily providing extra bandwidth to neighboring base stations. Independent of the problem nature, machine learning is used for analytical processing in Twitter [4]. Mugen Peng et al. [11] also provides some related survey of the applications of machine learning in wireless networks.

As can be seen in the figure 1 machine learning is generally classified into two types: supervised learning and unsupervised learning. The aim of supervised learning is to learn a mapping function from the training samples to the desired output, given a human-labeled data set. Unsupervised learning is to learn a function from unlabeled data without relying on human supervision [7].

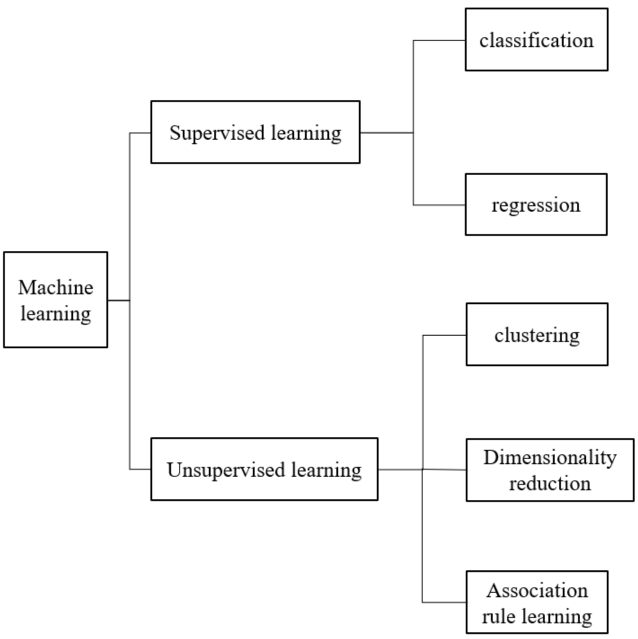


Fig. 1. The main classification of machine learning.

3.4 Clustering

Data analysis is a common method in modern scientific research. As the basic structure of data analysis, clustering plays an important role. As is shown in the Figure 1 that clustering is one technique of machine learning. Clustering is to divide similar data together and the specific division is not concerned with the label of the class. In other words, the similarities between data objects in the same cluster are as high as possible, while the disparities between data points in different clusters are likewise as high as feasible [12]. In addition, many cluster analysis tools have been created. Each clustering algorithm has its own advantages and disadvantages [13].

The study by Rui Xu et al. [12] offers two types of clustering: partitional clustering and hierarchical clustering. The basic idea of partitional clustering is that the center of each data point corresponds to the center of the related cluster, while hierarchical clustering algorithm aims to build a hierarchical relationship between data for clustering.

Data is classified into a specified number of clusters by it. Different clustering algorithm approaches usually generates different clusters, and even within the same algorithm, the identification of parameters may influence the final findings. Effective evaluation criteria and guidelines are critical for giving consumers trust in the clustering results produced by the algorithm. These evaluations should be objective, with no preference for any algorithm [14]. In this paper, the Sum of the Squared Errors (SSE) and Silhouette Coefficient are used as the evaluation standards for clustering.

3.4.1 K-means

K-means is a common clustering technique in data analysis that is commonly applied for grouping huge sets of data. In 1967, MacQueen firstly published a paper in which they described the k-means that was one of unsupervised learning algorithms [15].

This method aims at partitioning the given data set into k clusters. K is selected randomly as initial centroids and is determined in advance. Next step is calculating the distance between data points and each of the k centroids, and then divide each data point into the nearest cluster center. Initializing the clusters is done. A new center of mass is recalculated for each cluster and the data points are reassigned to the appropriate clusters. The above steps continue repeatedly until the position of the center of mass no longer changes or the set number of iterations is reached [16].

In this algorithm, Euclidean distance is used to calculate the distance of each data point X = $(x_1,x_2,\ x_3 \cdots x_i)$ and the centroid Y = $(y_1,y_2,\ y_3 \cdots y_i)$ and then determine the closet distance between them [15]. The Euclidean distance d (X, Y) is presented as follows:

$$\mathrm{d}\ (\mathrm{X},\ \mathrm{Y}) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_i - y_i)^2} \quad (1)$$

The most widely considered clustering criterion is the sum of the squared Euclidean distances (SSE) between each data object x_i and the centroid y_i of the subgroup C_k which consists of x_i.The SSE function is the sum of the distortion of the classes, and the distortion of each class is equal to the square of the distance from each point to the center of its class. The criterion function is obtained as follows [17]:

$$\mathrm{E} = \sum_{i=1}^{k} \sum_{x \in C_k} |x - x_i|^2 \quad (2)$$

The SSE is used to determine the value of k, which called the Elbow method [18]. As the value of k increases, the number of samples in each class decreases, so the samples are closer to their centers, and the average distortion reduces. As the value of k continues to grow, the improvement of the average distortion falls. In the process of increasing the k-value, when the distortion reaches a critical point, it improves dramatically before gradually declining, and this critical point is regarded the point with the best clustering performance. This means that the graph between SSE and k is shaped like an elbow, and the value of k, this elbow, represents the true cluster number for the data.

The input parameter of K-means is the number of clusters k and a dataset D. it can output a set of clusters. The main process is as below and figure 2:

(1) Randomly chose k from D as initial centroids.
(2) Calculate the distance between the data.
(3) Classify data objects into the nearest cluster.
(4) Recalculate distance and update the cluster centers based on the mean value.
(5) Repeat until no changing in the new centroid.

In this paper, the k-means algorithm is applied for Twitter location data, which is clustered into k groups. The result of it is analyzed and then to obtain the plan of cells ’deployment for optimizing wireless network in Barcelona.

3.4.2 DBSCAN

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) clustering is first proposed in 1996, which is the most commonly used clustering algorithm [14]. Moreover, it is density-based method which can detect the amount of clusters based on the data and to generate clusters for arbitrary shapes. This means it does not require any the number of clusters beforehand.

This technique divides regions with sufficient density into clusters and finds groups of arbitrary shapes in the presence of noise. A cluster is defined as the largest set of densely connected points. It only takes two parameters: the radius of the neighborhood around a point (Eps) and the number of objects contained in the neighborhood at least (MinPts) [14].

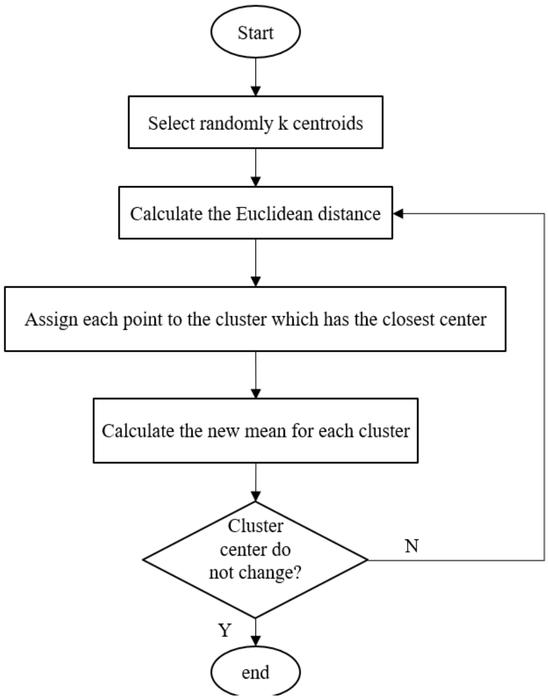


Fig.2. The flow chart of the k-means algorithm.

D is dataset and 'a' is an arbitrary object. The neighborhood of 'a' is defined as follow [19]:

$$N_{eps} = \{b \in D \, / dist(a,b) < Eps\} \tag{3}$$

If there are more neighborhoods of an object 'A' within that Eps range than Minpts, then it is marked as a core point. It can be described as follow:

$$N_{eps}(A) > Minpts \tag{4}$$

If it is surrounded by fewer objects within the given Eps range than Minpts, then the object is marked as noise or a border point.

The process of DBSCAN is shown as table 1.

Table 1. The Process of Dbscan

Algorithm: The DBSCAN clustering algorithm
Input: dataset D and Eps, Minpts.
Output: clusters of objects.
Process:
(1) select an object Parbitrarily.
(2) retrieve all objects from P.
(3) If $N_{eps}(P) > Minpts$, a cluster with P as a core point is formed.
(4) If P is a border object, then to visit the next object.
(5) Repeat until all objects have been visited.

Eps can be determined by drawing a k-distance graph [20]. It is obtained at the obvious inflection point in the k-distance graph. Most of the data will not be clustered if Eps is set too small, while most objects will be grouped into the same cluster if Eps is large. A rule of thumb is that MinPts is the value of k chosen in the k-distance graph. If MinPts obtains huge, it will lead to points in the same cluster being marked as outliers, and if MinPts is too small, a large number of core points will be found. Eps and MinPts can be adjusted appropriately if the results of empirical clustering are not satisfactory, and the most suitable parameters can be selected after several iterations of calculation and comparison.

Silhouette Coefficient is to identify whether DBSCAN clustering is well [21]. *c(i)* is intra-cluster dissimilarity, which is the average of the degree of dissimilarity from a point *i* to other points within the same cluster, reflecting cohesion. *d(i)* is inter-cluster dissimilarity, which is the minimum of the average dissimilarity from the object *i* to the other clusters, reflecting the degree of separation.

Silhouette Coefficient is obtained as follow:

$$s(i) = \frac{d(i)-c(i)}{max\,\{c(i),d(i)\}} \tag{5}$$

This formula can also be written as this:

$$s(i) = \begin{cases} 1-\frac{c(i)}{d(i)} & if\ c(i) < d(i) \\ 0 & if\ c(i) = d(i) \\ \frac{d(i)}{c(i)}-1 & if\ c(i) > d(i) \end{cases} \tag{6}$$

It easily can be seen from above definition that:

$$-1 \leq \mathrm{s(i)} \leq 1 \tag{7}$$

s(i) is close to 1, which implies that the point *i* is well clustered. When s(i) is approximately zero, then a(i) is equal to b(i). This indicates that the object i is on the boundary of two clusters and it is not clear which cluster this point should have been classified into. A different situation takes place when s(i) is nearly -1. It shows that the sample i is more deserving of classification into the other cluster.

4. Results & Evaluation

As presented above, clustering is crucial as data analysis tool and has been adopted in classification. In this section, there are two parts. We will firstly show the results used k-means algorithm and then discuss how the parameters influence results. The second part is to analyze the clusters applied DBSCAN technique and evaluate performances. We also give some suggestions for deploying cells of wireless network based on the result of them.

4.1 K-means Clustering

In the experiment, SSE is calculated and showed in figure 3. As discussed in the methodology, we chose the inflection point as the value of k. It is apparent from the figure 3 below that the SSE is turning at 3and then k is set to 3.

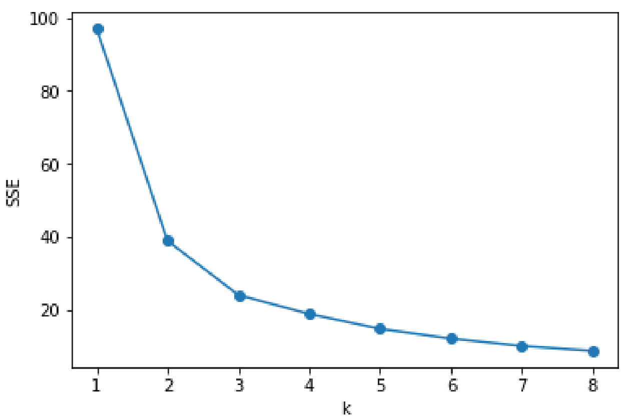


Fig.3. k-SSE graph.

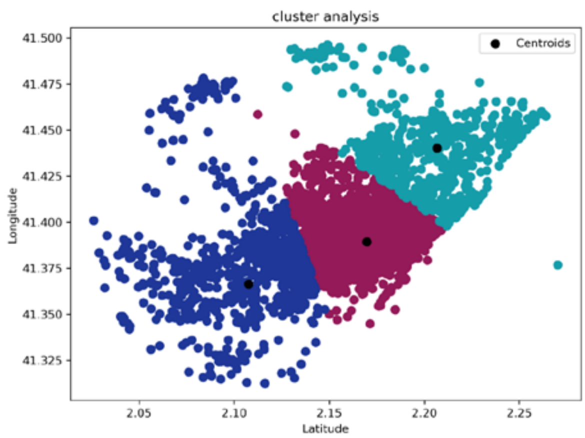


Fig.4. The result of k-means when k=3.

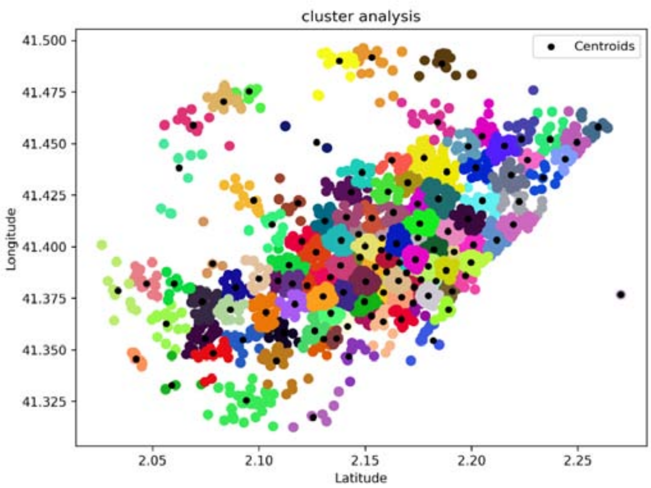
When k is 3, the clustering result applied the k-means technique is presented in Figure 4. The data in Barcelona is classified into three parts. These three centroids on Google Earth are residential areas where people use Twitter frequently.

The purpose of this experiment is to obtain the plan of the deployment of cells for wireless network. However, this result is not practical for deploying.

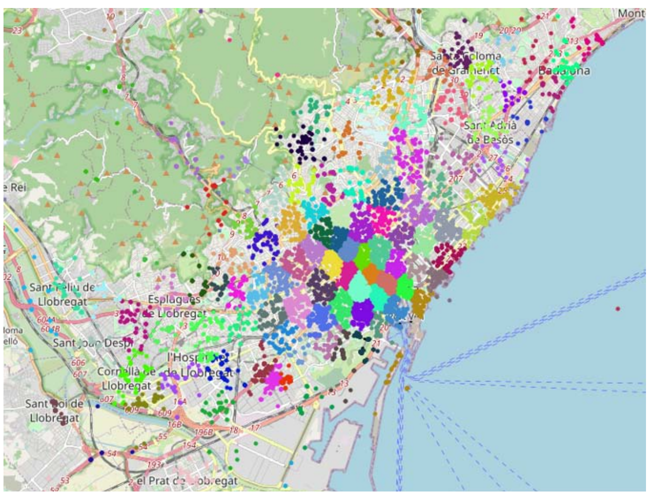
In practice, there are usually two special tiny areas within the area covered by large cells. One is the "blind spot", a shadow area caused by an obstacle in the propagation of the radio waves, where the quality of communication is seriously poor. The other is the "hotspot", an area of heavy traffic due to the uneven distribution of the spatial service load. Marco cells cover a large area with low power, while micro cells have a small coverage area with high power. Hence, microcells are mainly placed in "hotspots" within macrocells. The overlapping of cells of different radius and the presence of base stations with different transmitting powers in close proximity to each other give the entire wireless network a multi-layered structure. The reason why this result is not adopted is that the area covered by each cluster is too large to identify where to place the cells. Furthermore, according to this result, only macro cells can be placed in the corresponding positions, and it cannot show where to overlay micro cells. Therefore, it is difficult to optimize the wireless network and then we get another result based on actual area in Barcelona.

There are various types of cells depending on the radius. In this paper we use macro cells with radius of 1, 2.5 and 5 km respectively, and micro cell is also applied.

In cell-based mobile communication networks, a positive hexagonal wireless cell adjacency is usually used to form a faceted service area. According to the formula for the area of a hexagon, the coverage area of a macro cell with a radius of 1 km is $\frac{\sqrt{3}}{2}km^2$. It is well known that Barcelona covers an area of 101.9 square kilometers. Therefore, the cluster number is about 118, which is equal to the area of Barcelona divided by the area covered by one macro cell. In other word, macro cells with 1 km need to be placed 118 in Barcelona and there are 118 micro cells installed in each centroid. The result is presented as figure 5.



(a) The result of the k-means when k =118

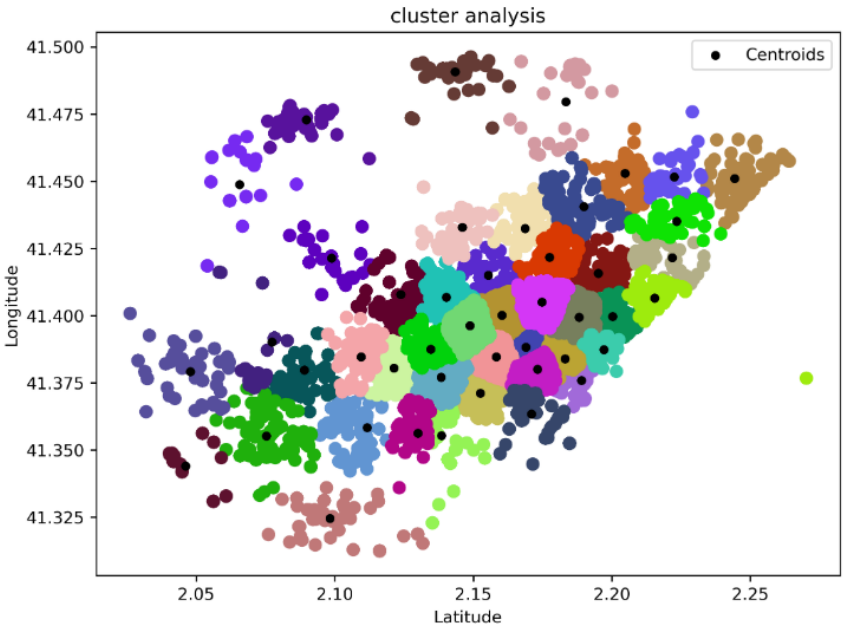


(b) The result of the k-means shown on the map when k =118

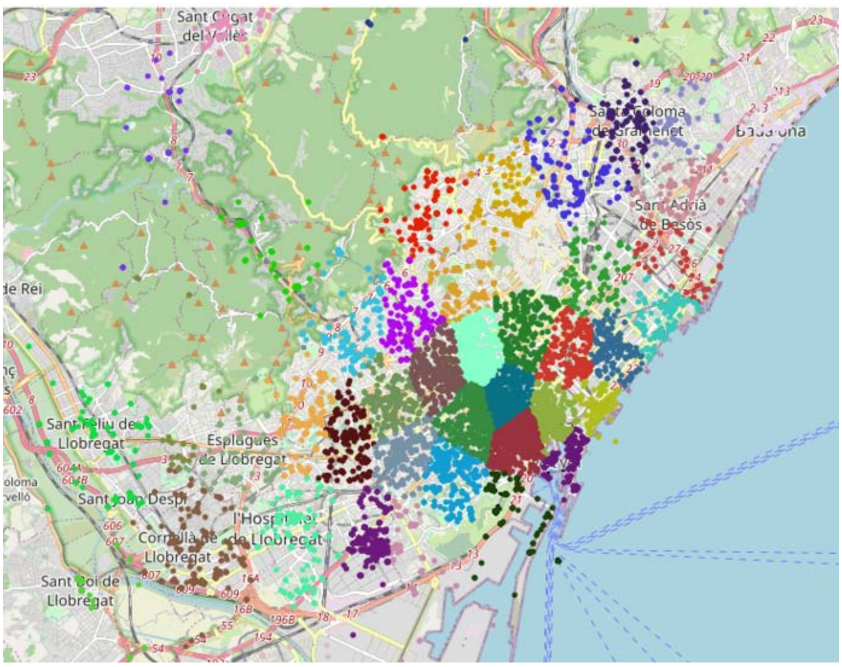
Fig.5. (a)(b) The cluster result of k-means when k =118.

From the graph above we can see that the data in the middle are more aggregated and clustered clearly. On the other hand, Peripheral data is too discrete, the plan for them to deploy is not very suitable. This means that creating cells outside the city where these data are scattered could waste resources because no one uses the cells in some areas.

The following figure 6 and 7 is the cluster result of macro cells with radius of 2.5 and 5 km respectively. The process of these two cells is similar to installing the macro cells with radius of 1 km. The coverage area of a macro cell with a radius of 2.5 km and 5 km is $\frac{5\sqrt{3}}{4}km^2$ and $\frac{5\sqrt{3}}{2}km^2$ respectively.Therefore, the cluster number is about 45 and 24 based on the area covered in Barcelona. Centroids are considered as hotspots and then micro cell is placed at each cluster center.

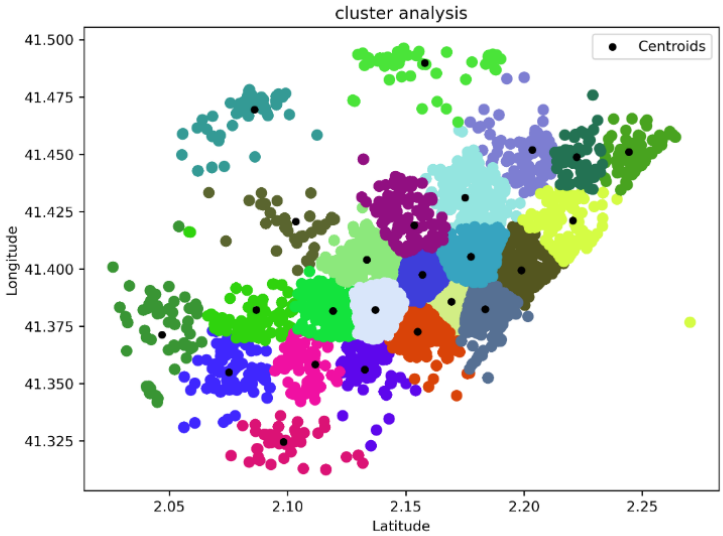


(a) The result of the k-means when k =45

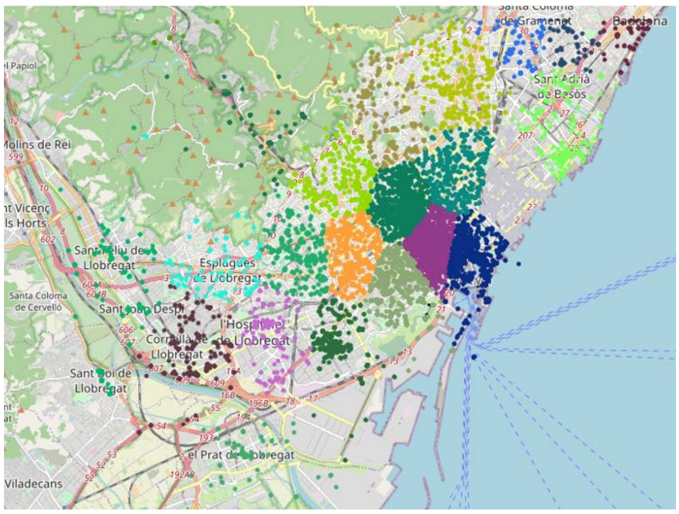


(b) The result of the k-means shown on the map when k =45

Fig.6. (a)(b) The cluster result of the k-means when k =45.



(a) The result of the k-means when k =24



(b) The result of the k-means shown on the map when k =24

Fig.7. (a)(b) The cluster result of the k-means when k =24.

The overall result of the k-means for optimizing wireless network is showed as table 2.

Table 2. Plan for Wireless Network

	Coverage	The number of cells
	1 km	118
Macro cell	2.5 km	45
	5 km	24
Micro cell	100 m - 1 km	centroids

The main weakness with this algorithm is that it is not suitable for peripheral classes to deploy. Another problem with this approach is that the cluster center is selected randomly. In summary, this result is local optimal solution.

DBSCAN Clustering

DBSCAN algorithm requires the determination of only two parameters: the radius of the neighborhood around a point (Eps) and the minimum number of objects required (Minpts). The k-distance graph is obtained as figure 8.

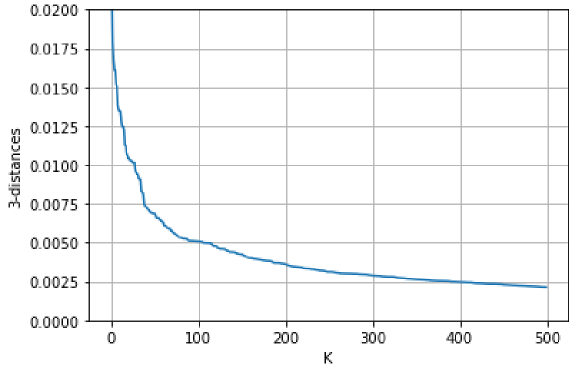


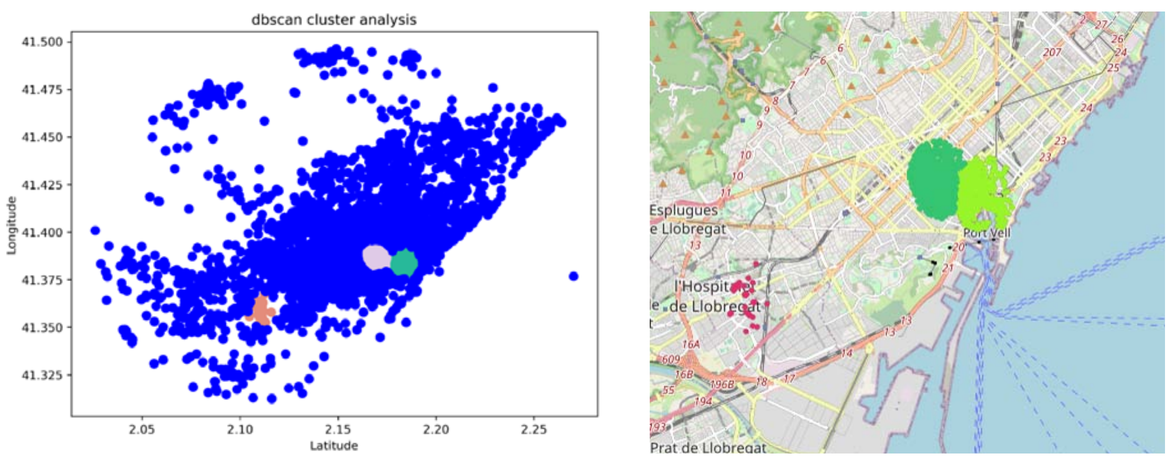
Fig.8. The k-distance graph.

Eps is regarded as the threshold point is the first point in the first "valley" of the k-distance graph. For Knee point detection, the KneeLocator module in python is used and it prints 0.004667. As can be seen from the data in figure that the value is 0.005. The result of them is very similar and 0.004667 is chosen. The table 3 below illustrates the Silhouette coefficient is 0.960517 when Eps is 0.004667 and Minpts is 2000. This coefficient is close to 1, which implies that objects are well clustered.

Table 3. Parameters for DBSCAN

Eps	Minpts	Silhouette coefficent	Cluster number
0.004667	5	0.196454	38
0.004667	10	0.202195	29
0.004667	50	0.372666	21
0.004667	100	0.461426	17
0.004667	200	0.530824	13
0.004667	500	0.425580	12
0.004667	1000	0.447651	7
0.004667	1500	0.599463	4
0.004667	2000	0.960517	3

The data can be categorized into three sub-groups. Figure 9 (a) provides the clustering results of DBSCAN algorithm and the blue region represents noise. Figure 9(b) shows the cluster without noise. Compared with map of Barcelona by district, the leftmost area is located at I' Hospital de Llobregat, the rightmost region is about Ciutat Vella and the rest one is Eixample. The following paragraphs describe the deployment of cells for these three districts.



(a) The result of DBSCAN (b) The result of DBSCAN without noises shown on the map

Fig.9. (a)(b) The cluster result of DBSCAN.

The light green region is Ciutat Vella, which covers 5km^2.The coverage area of a macro cell with a radius of 1 km, 2.5km and 5 km is $\frac{\sqrt{3}}{2}$km^2, $\frac{5\sqrt{3}}{4}$km^2 and $\frac{5\sqrt{3}}{2}$km^2 respectively. According to this, the number of macro cells is 6, 3 and 2 respectively. In addition, we also apply micro cells with radius of 100 m, 300 m and 500 m respectively. They cover $\frac{\sqrt{3}}{20}$m^2, $\frac{3\sqrt{3}}{20}$m^2 and $\frac{\sqrt{3}}{4}$m^2 respectively. Therefore, the number of micro cells to be set is 58, 20 and 12. The general plan for cells is summarized in table 4.

Table 4. PLAN for Ciutat Vella

	Coverage	The number of cells
Macro cell	1 km	6
	2.5 km	3
	5 km	2
Micro cell	100 m	58
	300 m	20
	500 m	12

Similarly, the overall plan for Eixample to place cells is presented in table 5 based on its area of 7.485km^2.

Table 5. PLAN for Eixample

	Coverage	The number of cells
Macro cell	1 km	9
	2.5 km	4
	5 km	2
Micro cell	100 m	87
	300 m	29
	500 m	18

The data in I' Hospital de Llobregat does not cover the whole area, so it is not possible to use the area to calculate the quantity of cells.As shown in Figure 9(b), the data is almost divided into two regions and there are about 20 points in this purple area. Therefore, the plan for this region to deploy cells is shown as tableVI.

Table 6. PLAN for I' Hospital de Llobregat

	Coverage	The number of cells
Macro cell	1-2.5 km	2
Micro cell	100 m - 1 km	20

5. Conclusions

The aim of this paper is to analyze the distribution of hotspots in Barcelona based on Twitter data and then to optimize wireless network.

In this paper, we first have presented the reason for choosing social media data. After describing the methodology including machine learning, we have illustrated clustering algorithms, which are attractive for data classification. There are two clustering techniques: K-means and DBSCAN applied for data analysis. Finally, we have shown suggestions about the deployment of cells based on the clustering results used on these two algorithms. K-means set three k values according to the coverage of macro cells and then give suggestions about how many cells can be placed. Another algorithm is DBSCAN. Our results presented that this algorithm performed well when Eps is 0.004667 and Minpts is 2000. Three districts were classified and plan for install cells also showed based on clustering results. The results of our experiment can help operators develop the deployment of cells and then improve wireless network communication quality.

6. Future Work

Future research will consider the following issues. For the result of k-means algorithm, peripheral data is too discrete, the plan for them to deploy is not very suitable. For data that clusters well in the center, this algorithm is used to analyze it, and for data that is more dispersed in the periphery, other algorithms are used instead. For k-means algorithm, the initial centroids have a significant impact on the quality of the final clustering results, but the initial centroids are chosen arbitrarily. The K-Means++ algorithm is an optimization of the K-Means method of randomly initializing the center. Moreover, this algorithm is sensitive to noise and abnormal points. In the future, LOF algorithm can be used to improve this, which can remove outliers and then reduce the impact of outliers and noise on the clustering.

For the result of DBSCAN algorithm, the deployment of cells for I' Hospital de Llobregat is roughly estimated, so other ways are considered to accurately deploy cells. For DBSCAN algorithm, although this algorithm only needs to select two parameters, people need to modify them repeatedly and compute whether the Silhouette coefficient is near to 1. Simon Fong et al. [19] summarized enhanced DBSCAN algorithm to get the efficient clustering results. Using these methods can improve clustering efficiency and reduce computational complexity.

References

[1] N. Khan *et al.*, "Big Data: Survey, Technologies, Opportunities, and Challenges," *Sci. World J.*, vol. 2014, p. 712826, 2014, doi: 10.1155/2014/712826.

[2] S. Bi, R. Zhang, Z. Ding, and S. Cui, "Wireless communications in the era of big data," *IEEE Commun. Mag.*, vol. 53, no. 10, pp. 190–199, 2015, doi: 10.1109/MCOM.2015.7295483.

[3] M. Ballings, D. Van den Poel, and M. Bogaert, "Social media optimization: Identifying an optimal strategy for increasing network size on Facebook," *Omega (United Kingdom)*, vol. 59, pp. 15–25, 2016, doi: 10.1016/j.omega.2015.04.017.

[4] D. G. Costa, C. Duran-Faundez, D. C. Andrade, J. B. Rocha-Junior, and J. P. J. Peixoto, "TwitterSensing: An event-based approach for wireless sensor networks optimization exploiting social media in smart city applications," *Sensors (Switzerland)*, vol. 18, no. 4, 2018, doi: 10.3390/s18041080.

[5] M. I. Jordan and T. M. Mitchell, "Machine learning: Trends, perspectives, and prospects," vol. 349, no. 6245, 2015.

[6] W. Qi, R. Procter, J. Zhang, and W. Guo, "Mapping consumer sentiment toward wireless services using geospatial twitter data," *IEEE Access*, vol. 7, pp. 113726–113739, 2019, doi: 10.1109/ACCESS.2019.2935200.

[7] Y. Liu, S. Bi, Z. Shi, and L. Hanzo, "When Machine Learning Meets Big Data: A Wireless Communication Perspective," *IEEE Veh. Technol. Mag.*, vol. 15, no. 1, pp. 63–72, 2020, doi: 10.1109/MVT.2019.2953857.

[8] Y. Ren, "Python Machine Learning : Machine Learning and Deep Learning With Python ," *Int. J. Knowledge-Based Organ.*, vol. 11, no. 1, pp. 67–70, 2021.

[9] N. A. Ghani, S. Hamid, I. A. Targio Hashem, and E. Ahmed, "Social media big data analytics: A survey," *Comput. Human Behav.*, vol. 101, no. December 2017, pp. 417–428, 2019, doi: 10.1016/j.chb.2018.08.039.

[10] B. Yang, W. Guo, B. Chen, G. Yang, and J. Zhang, "Estimating Mobile Traffic Demand Using Twitter," *IEEE Wirel. Commun. Lett.*, vol. 5, no. 4, pp. 380–383, 2016, doi: 10.1109/LWC.2016.2561924.

[11] Y. Sun, M. Peng, Y. Zhou, Y. Huang, and S. Mao, "Application of Machine Learning in Wireless Networks: Key Techniques and Open Issues," *IEEE Commun. Surv. Tutorials*, vol. 21, no. 4, pp. 302–3108, 2019, doi: 10.1109/COMST.2019.2924243.

[12] J. Yu, H. Huang, and S. Tian, "Cluster validity and stability of clustering algorithms," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 3138, no. 3, pp. 957–965, 2004, doi: 10.1007/978-3-540-27868-9_105.

[13] D. Xu and Y. Tian, "A Comprehensive Survey of Clustering Algorithms," *Ann. Data Sci.*, vol. 2, no. 2, pp. 165–193, 2015, doi: 10.1007/s40745-015-0040-1.

[14] T. Ali, S. Asghar, and N. A. Sajid, "Critical analysis of DBSCAN variations," *2010 Int. Conf. Inf. Emerg. Technol.*

ICIET 2010, 2010, doi: 10.1109/ICIET.2010.5625720.

[15] N. Shi, X. Liu, and Y. Guan, “Research on k-means clustering algorithm: An improved k-means clustering algorithm,” *3rd Int. Symp. Intell. Inf. Technol. Secur. Informatics, IITSI 2010*, pp. 63–67, 2010, doi: 10.1109/IITSI.2010.74.

[16] C. Zhang and S. Xia, “K-means clustering algorithm with improved initial center,” *Proc. - 2009 2nd Int. Work. Knowl. Discov. Data Mining, WKKD 2009*, vol. 1, no. 2, pp. 790–792, 2009, doi: 10.1109/WKDD.2009.210.

[17] A. Likas, N. Vlassis, and J. J. Verbeek, “The global k-means clustering algorithm,” *Pattern Recognit.*, vol. 36, no. 2, pp. 451–461, 2003, doi: 10.1016/S0031-3203(02)00060-2.

[18] J. Guo, Y. Li, M. Hou, S. Han, and J. Ren, “Recognition of daily activities of two residents in a smart home based on time clustering,” *Sensors (Switzerland)*, vol. 20, no. 5, pp. 1–15, 2020, doi: 10.3390/s20051457.

[19] S. Fong, S. U. Rehman, K. Aziz, and I. Science, “DBSCAN : Past, Present and Future,” pp. 232–238, 2014.

[20] M. Daszykowski and B. Walczak, “Density-Based Clustering Methods,” *Compr. Chemom.*, vol. 2, pp. 635–654, 2009, doi: 10.1016/B978-044452701-1.00067-3.

[21] P. J. Rousseeuw, “Silhouettes: A graphical aid to the interpretation and validation of cluster analysis,” *J. Comput. Appl. Math.*, vol. 20, no. C, pp. 53–65, 1987, doi: 10.1016/0377-0427(87)90125-7.