Chapter

Practical Application Using the Clustering Algorithm

Yoosoo Oh and Seonghee Min

Abstract

This chapter will survey the clustering algorithm that is unsupervised learning among data mining and machine learning techniques. The most popular clustering algorithm is the K-means clustering algorithm; It can represent a cluster of data. The K-means clustering algorithm is an essential factor in finding an appropriate K value for distributing the training dataset. It is common to find this value experimentally. Also, it can use the elbow method, which is a heuristic approach used in determining the number of clusters. One of the present clusterings applied studies is the particulate matter concentration clustering algorithm for particulate matter distribution estimation. This algorithm divides the area of the center that the fine dust distribution using K-means clustering. It then finds the coordinates of the optimal point according to the distribution of the particulate matter values. The training dataset is the latitude, longitude of the observatory, and PM10 value obtained from the AirKorea website provided by the Korea Environment Corporation. This study performed the K-means clustering algorithm to cluster feature datasets. Furthermore, it showed an experiment on the K values to represent the cluster better. It performed clustering by changing K values from 10 to 23. Then it generated 16 labels divided into 16 cities in Korea and compared them to the clustering result. Visualizing them on the actual map confirmed whether the clusters of each city were evenly bound. Moreover, it figures out the cluster center to find the observatory location representing particulate matter distribution.

Keywords: clustering, machine learning, data mining, K-means clustering, particulate matter

1. Introduction

This chapter introduces the data mining and the clustering algorithm, which is unsupervised learning among machine learning techniques. In this chapter, we analyze the performed clustering application research that used the air pollution concentration data. It has been a problem recently. The most popular algorithm among the clustering is the K-means clustering algorithm; it represents a data cluster. It is an essential factor that finds an appropriate K value for the distribution of the training dataset. Commonly, we determine the K value experimentally, and at this point, we can set the value using the elbow technique.

One example of clustering application studies is the air pollution concentration clustering algorithm. Air pollution is a substance that causes respiratory diseases and cancer, and the WHO reported the severity of the particulate matter [1–3].
The Korean government has also started providing particulate matter and air pollution information since 2004. On the AirKorea website, we can obtain air pollution information measured at 353 observatories in real-time [4].

Currently, observatories of air pollution in Korea are mainly located in Seoul and Gyeonggi-do, so it is challenging to know accurate air pollution values in local small towns without observatories. Therefore, in this chapter, we study the clustering method for air pollution observatory according to the air pollution concentration. We first split the air pollution-centered regions that can predict the distribution of air pollution by using K-means clustering. Then, we find the optimal station location according to the distribution of air pollution concentrations. Based on the optimal location, we divide the territory of the Korean.

We collect air pollution data in April 2020 and label air pollution monitoring stations through clustering algorithms for this clustering study. Based on the cluster center point, we can apply the Voronoi algorithm to divide the territory of Korea. With this method, we can classify air pollution areas by considering the concentration distribution of air pollution, unlike traditional administrative districts. Furthermore, this method can help know the air pollution distribution in the shaded area without air pollution [5, 6].

2. Related works

In this section, we analyzed related studies to predict the concentration of fine dust [7–12]. The related studies use air pollution data and meteorological data together. In particular, the accuracy of prediction is high when weather data such as temperature and wind speed are used rather than air pollution data [7]. Traditionally, the studies predict the concentration of fine dust through machine learning methods such as linear regression or support vector regression. However, these methods are challenging to consider the spatiotemporal correlation [8]. Therefore, it focuses on improving prediction accuracy by using deep learning [9–12]. There are four distinct seasons in Korea depending on the air mass, so there is a significant difference in the concentration of fine dust by season. Therefore, we must be considered the relationship between location and time.

Joun et al. predicted the concentration of fine dust using the MLR, SVR, ARIMA, and ARIMAX [11]. In this paper, the training datasets are air pollution data (NO2, SO2, CO, O3, PM10) and meteorological data (temperature, precipitation, wind speed). They confirmed that time, location, NO2, CO, O3, SO2, maximum temperature, precipitation, and maximum wind speed were significant variables using multiple linear regression analysis. In addition, they used multiple linear regression and support vector regression to predict fine dust distribution. The prediction accuracy was higher in the artificial neural network than in the multiple support vector regression. If the PM10 concentration increased above 100, the support vector regression was exceptionally high. They perform experiments using ARIMA and ARIMAX to analyze the factors of time according to the location. As a result, there was a difference in the learning accuracy according to the location of the experimental data. Furthermore, the accuracy was higher in using the air quality factor and the meteorological factor than using only the time variable.

DOI: http://dx.doi.org/10.5772/intechopen.99314

Cho et al. designed a predictive model through multiple linear regressions and artificial neural networks and performed the fine-dust prediction [12]. They collected the training data, air pollution data (NO2, SO2, CO, O3, PM10), and meteorological data (temperature, humidity, wind speed, wind direction).

As a result of analyzing the errors by performing prediction, the accuracy of the prediction model using artificial neural networks was better overall than that of multiple linear regression. As the result of the experiment by changing the hidden layers of the artificial neural network, the performance of the multi-layer perceptron was better when there were three hidden layers.

3. Air pollution data mining

The algorithm for air pollution concentration clustering performs clustering through observatories' location and measurement values. We can utilize air pollution information in AirKorea provided by the Korea Environment Corporation. The values of NO2, SO2, CO, O3, PM10, PM25 are upload every hour [13].

In this chapter, we use air pollution information measured in April 2020. First, we download national data on a website to create a feature dataset. And then, we should be converted the address of the observatory to latitude and longitude coordinates of the WGS84 coordinate system. We used Kakao Map API [14]. **Figure 1** shows the used dataset. For example, the 1 row is air pollution code is 111121, a date is April 1, 2020, and the location is the nearby city hall of Seoul.

The feature dataset used for air pollution clustering is latitude and longitude, NO2, SO2, CO, O3, PM10, PM25. We calculate an average observatory data to make one-day data into 1-hour data. Also, we filled in the missing values for each station by obtaining the value of the other stations closest to it. Pseudocode 1 is performing this process.

```
SET myData to READ(fileName)
SET feature to ['latitude', 'longitude', 'PM10', 'SO2', 'CO', 'O3', 'NO2', 'PM25']
IF type of 'date' in myData is string THEN
   convert datetime type to string type of 'date'
ELSE
   PASS
ENDIF
SET group to 'feature data' by 'date' in 'station code'
CALCULATE AVERAGE 'feature data' by group
```

Pseudocode 1.
The process of making the dataset by the day.

4. Air pollution area division method

In this section, we perform location labeling through K-means clustering using created air pollution dataset. The K-means clustering algorithm can be partitioning all data by defining the number of clusters and obtaining the center point of clusters. We used the Scikit-learn python library [15].

Station Code	Date	SO2	CO	O3	NO2	PM10	PM25	Latitude	Longitude
111121	2020-04-01 00:00:00	0.003435	0.443478	0.039565	0.015826	61.0869565	34	37.56455491	126.975615
111122	2020-04-01 00:00:00	0.003609	0.508696	0.035304	0.023826	71.3478261	32.9565217	37.55483769	126.9717341
111123	2020-04-01 00:00:00	0.004619	0.5	0.037095	0.017571	61.8571429	29.7619048	37.57206081	127.0050305
111124	2020-04-01 00:00:00	0.004261	0.578261	0.044304	0.023864	70.0869565	37.2608696	37.56863712	126.9981844
111125	2020-04-01 00:00:00	0.003783	0.473913	0.044783	0.01813	73	32.6956522	37.5709184	126.9965543
111131	2020-04-01 00:00:00	0.003524	0.504762	0.040952	0.020524	71	29.8095238	37.54043943	127.0042997
111141	2020-04-01 00:00:00	0.003217	0.513043	0.040652	0.018913	78.9130435	33.5454545	37.54655201	127.0921031
111142	2020-04-01 00:00:00	0.003783	0.46087	0.044565	0.02187	63.6521739	30.3043478	37.54307024	127.0417993
111143	2020-04-01 00:00:00	0.004087	0.430435	0.030522	0.039522	78.5217391	27.2173913	37.53895452	127.0416703
111151	2020-04-01 00:00:00	0.003632	0.510526	0.043842	0.021842	70	31.2631579	37.58491128	127.0940388
111152	2020-04-01 00:00:00	0.003913	0.513043	0.044391	0.020826	61.9130435	28.0434783	37.57641471	127.0283873
111154	2020-04-01 00:00:00	0.003957	0.552174	0.042217	0.031304	66.4347826	31.7391304	37.58045345	127.0443697

Figure 1.
The air pollution dataset in April 2020.

Because it has to be performed clustering by calculating the distance between each data, we normalize to a value between 0 and 1 using MinMaxScaling [15]. Moreover, we perform the K-means clustering using normalized data. Eq. (1) shows the dataset that maximizes the degree of cohesion within each set. The distance between it in each cluster is measured using the Euclidean distance. The clustering algorithm ends when the average value of this distance no longer changes.

$$\arg\min \sum_{i=1}^{k} \sum_{x \in S_i} \|x - \mu_i\|^2 \tag{1}$$

We calculate the inertia value to find the appropriate K value for K-means Clustering. The inertia value is the sum of the distances between clusters at each center point after clustering. **Figure 2** shows the inertia value according to the K. The optimal k value is where the inertia value decreases rapidly, and the change is not significant. However, it is difficult to determine the optimal k value in this graph. Therefore, we set the k value to 16, focusing on dividing the whole country into 16 provinces.

```
SET myData to READ(fileName)
SET feature to ['latitude', 'longitude', 'PM10', 'SO2', 'CO', 'O3', 'NO2', 'PM25']
FOR day to 31 DO
  GET 'feature' data in myData on 'day'
  CALL MinMaxScaler()
  CALL k-MeansClustering(k is 16)
```

Pseudocode 2.
The process of performing scaling and clustering.

Pseudocode 2 is the source code that loads April data, performs the scaling and clustering. Also, **Figure 3** presents the coordinates of the center point of each cluster as a result of performing clustering based on the air pollution data for a month. We use the Folium python library to show this map [16]. The marker of the same color is the cluster's point divided into 16 in the cluster for the day. Also, **Table 1** is a comparison of 16 administrative district labels and clustering results. For example, the 0 label is the Gangwon-do area, and 11, 12, 15 labels contain the twelve air pollution stations in this district.

DOI: http://dx.doi.org/10.5772/intechopen.99314

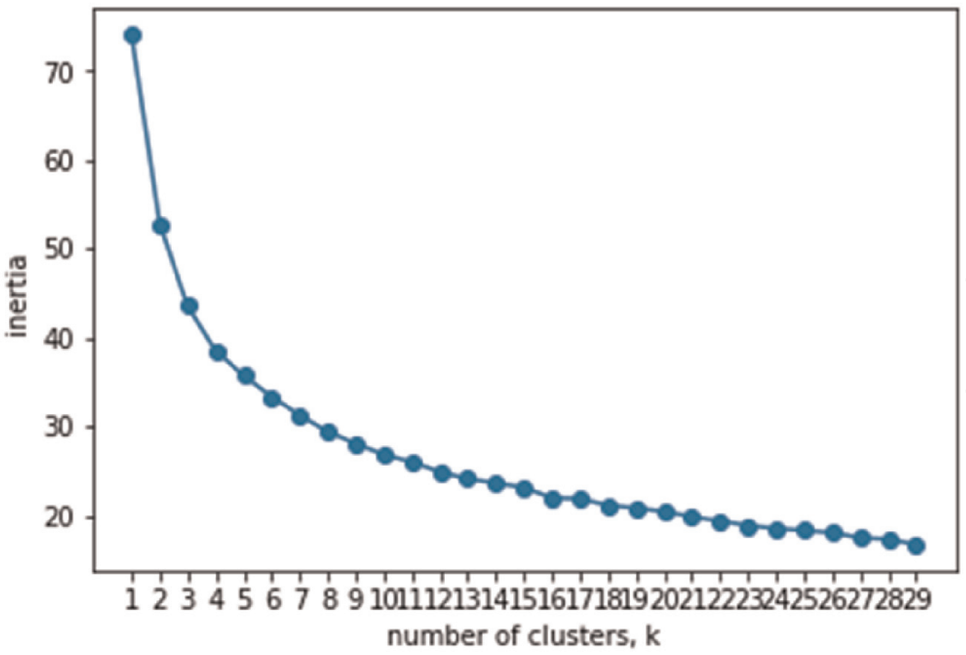

Figure 2.
The inertia value according to the K.

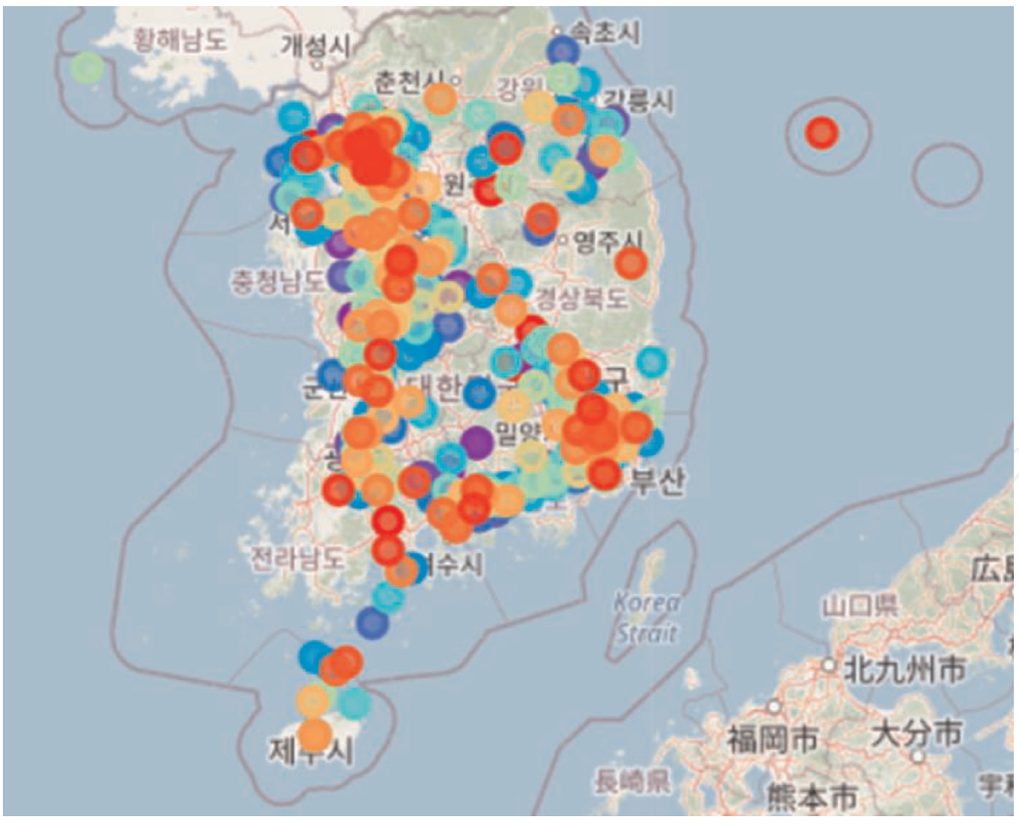

Figure 3.
The coordinates of the center point of each cluster for a month.

To determine the coordinates of the 16 cluster's center point, we perform the K-means clustering again. **Figure 4** is the visualization of the 16 center points on a map to divide regions. As a result of performing clustering, it is more closely

Administrative District	Clustering Label															
	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Gangwon-do												3	4			5
Gyeonggi-do	22			11			38						2		7	
Gyeongsangnam-do		1	6						1					9		
Gyeongsangbuk-do		5						5	1		1	5		3		
Gwangju			1		8											
Daegu		9						3			1			1		
Deajeon						6			3						1	
Busan								7			4			10		
Seoul	8			10			21									
Ulsan								5			9			2		
Incheon	3			12												
Jeollanam-do			8		8					3						
Jeollabuk-do					3	2			14							
Jeju-do										5						
Chungcheongnam-do	1					15			1						6	
Chungcheongbuk-do						6							6		1	1

Table 1.
Number of stations in each cluster by administrative district.

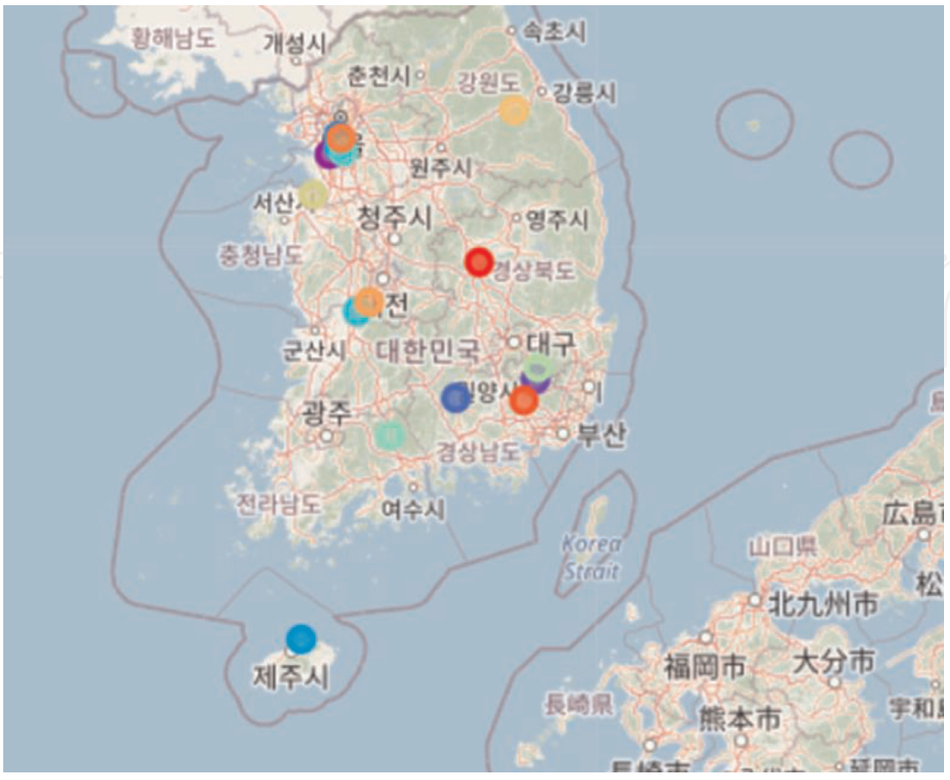

Figure 4.
The visualization of the 16 center points on a map to divide regions.

DOI: http://dx.doi.org/10.5772/intechopen.99314

distributed in the Seoul cluster center, Incheon, and Gyeonggi-do than other regions. It is because many air pollution monitoring stations are mainly distributed in the metropolitan area in Korea.

Figure 5 shows the results of classifying air pollution monitoring stations by calculating the distance to each station from the obtained 16 center coordinates. Points on the map are the location of the air pollution monitoring station. In this case, we calculated the Euclidean distance using latitude and longitude.

Also, **Figure 6** visualizes the convex hull polygon by connecting the outermost point of the classified measurement stations as a line [17]. This method has the advantage of accurately classifying even if the distance between each point is close because classification is performed based on the location of the stations. However, in an area without an observatory, it is a shaded area, and the distribution of air pollution cannot be measured.

This chapter found cluster's center points using the location and concentration of air pollution monitoring stations to divide air pollution areas that can reflect data distribution. The stations are classified based on the center coordinates, and air pollution areas are divided using the Convexhull polygon. However, there was a problem that the classified air pollution areas did not include areas without air pollution monitoring stations.

Therefore, we use the Voronoi algorithm to include areas without measurement stations [18]. Also, it can classify areas based on the center point of the cluster. The Voronoi algorithm is to get a line segment that can divide the distance between neighboring points into two and obtain a polygon with the intersection of each line segment as a vertex. **Figure 7** shows the divided regions using the Voronoi algorithm. The dots represent the centers of classified clusters. The method used in the Voronoi

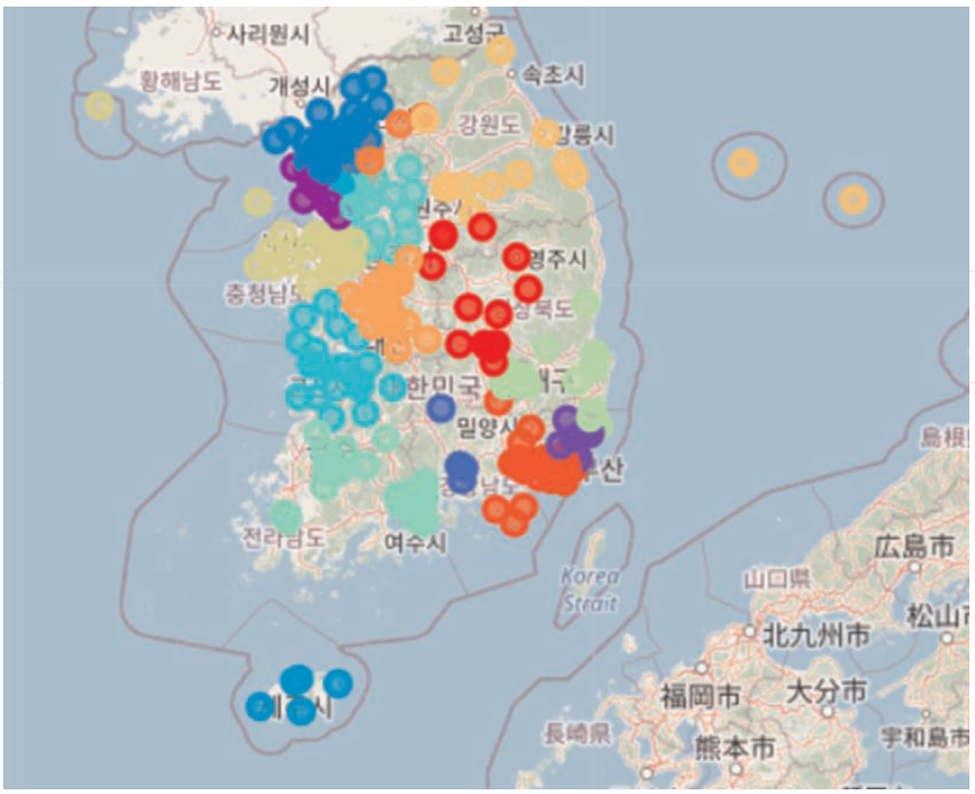

Figure 5.
The results of classifying air pollution monitoring stations by cluster.

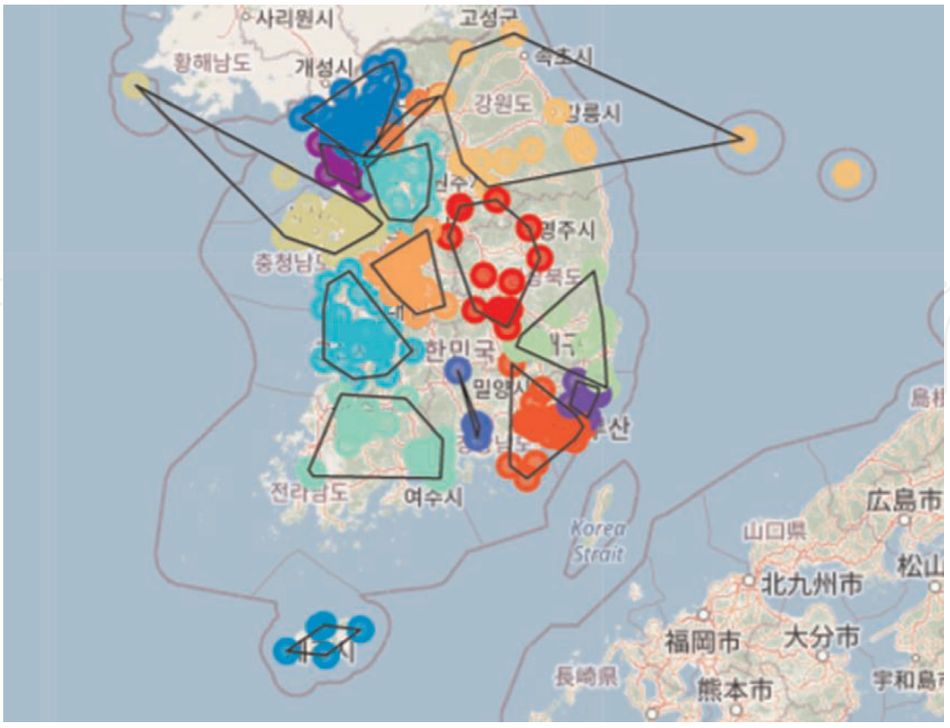

Figure 6.
The result using the convex hull polygon algorithm.

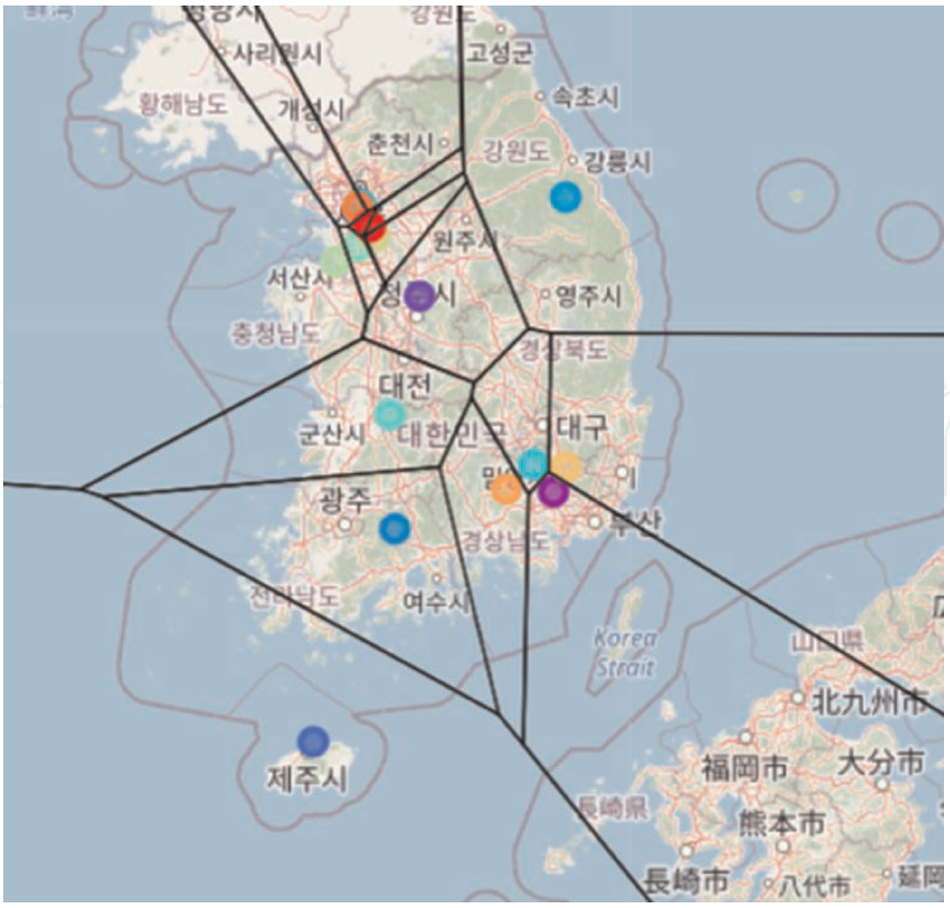

Figure 7.
The result using the Voronoi algorithm.

DOI: http://dx.doi.org/10.5772/intechopen.99314

algorithm is the Euclidean calculation method. Unlike the convex hull method in **Figure 6**, the Voronoi algorithm's classification method can divide regions without shadowed areas of the Korean Peninsula.

We compare the existing administrative districts in Korea [19], the regional classification method using the convex hull method, and the Voronoi algorithm. Existing administrative districts are classified according to the criteria defined in the Administrative District Practice Manual. Also, the convex hull method divided the area into classified air pollution measurement stations. The Voronoi algorithm classifies regions using the distance value based on the center point of the cluster. Air pollution concentrations were not reflected in existing administrative districts, but the convex hull method and Voronoi algorithm can classify regions. Finally, in the convex hull method, the area without a measuring station is shaded, unlike the existing administrative area and Voronoi algorithm. Comprehensively, the Voronoi algorithm can classify the region by reflecting the air pollution concentration without the shaded area.

5. Conclusion

In this chapter, we collected the data of air pollution stations in Korea and used K-means clustering to learn about data mining and machine learning algorithms. We divide air pollution areas to predict the distribution of air pollution using air pollution concentration clustering. The training dataset is latitude, longitude, NO2, SO2, CO, O3, PM10, PM25, with air pollution data for one month in April 2020. We use the collected dataset and classify air pollution monitoring stations. Based on the central coordinates of the cluster, the areas of the Korean territory were classified through the Voronoi algorithm. Finally, we confirmed that the proposed air pollution area could be classified by considering the distribution of air pollution, unlike traditional administrative districts. Moreover, the proposed area can help understand the distribution of air pollution in the shaded areas that do not have air pollution stations.

Acknowledgements

This research was supported by the Daegu University, 2018.

Author details

Yoosoo Oh* and Seonghee Min
Daegu University, Gyeongsan-si, Republic of Korea

*Address all correspondence to: yoosoo.oh@daegu.ac.kr

References

[1] WHO, Air pollution, May 2018, Available from: https://www.who.int/news-room/fact-sheets/detail/ambient-(outdoor)-air-quality-and-health/ [Accessed: 2021-06-01]

[2] Hänninen O. O. WHO Guidelines for Indoor Air Quality: Dampness and Mold. In Fundamentals of mold growth in indoor environments and strategies for healthy living. Wageningen Academic Publishers, Wageningen. 2011. p. 277-302.

[3] World Health Organization, WHO air quality guidelines global update, report on a working group meeting, Bonn, Germany, 18–20 October, 2005.

[4] Air Korea, Available from: http://www.airkorea.or.kr/ [Accessed: 2021-06-01]

[5] Min S, Oh, Y. A Study of Particulate Matter Clustering for PM10 Distribution Prediction, In: Proceedings of the International Symposium on Innovation in Information Technology and Applications (2019 ISIITA); 11-13 February 2019; Okinawa. p. 53-56.

[6] Min S, Oh Y. A study of particulate matter area division using PM10 data clustering: Focusing on the case of Korean particulate matter observatory. Journal of Adv Research in Dynamical and Control Systems. 2019;11.12: 959-965. DOI:10.5373/JARDCS/V11SP12/20193300

[7] Munir S, Habeebullah TM, Seroji AR, Morsy EA, Mohammed AM, Saud, WA, Awad AH. Modeling particulate matter concentrations in Makkah, applying a statistical modeling approach. Aerosol Air Quality Research. 2013;13.3:901-910.

[8] Li X, Peng L, Hu Y, Shao J, Chi T. Deep learning architecture for air quality predictions. Environmental Science and Pollution Research. 2016;23.22: 22408-22417.

[9] Freeman BS, Taylor G, Gharabaghi B, Thé J. Forecasting air quality time series using deep learning. Journal of the Air and Waste Management Association. 2018;68.8:866-886.

[10] Qi Z, Wang T, Song G, Hu W, Li X, Zhang Z. Deep air learning: Interpolation, prediction, and feature analysis of fine-grained air quality. IEEE Transactions on Knowledge and Data Engineering. 2018;30.12: 2285-2297.

[11] Joun S, Choi J, Bae J. Performance Comparison of Algorithms for the Prediction of Fine Dust Concentration. In: Proceedings of Korea Software Congress 2017, 8-10 February 2019; Pyeong Chang. p. 775-777.

[12] Cho K, Jung Y, Kang C, Oh C. Conformity assessment of machine learning algorithm for particulate matter prediction. Journal of the Korea Institute of Information and Communication Engineering. 2019;23.1:20-26.

[13] AirKorea, Available from: http://www.airkorea.or.kr/ [Accessed: 2021-06-01]

[14] Kakao Map API, Available from: https://apis.map.kakao.com/ [Accessed: 2021-06-01]

[15] Scikit-learn, Available from: https://scikit-learn.org/ [Accessed: 2021-06-01]

[16] Folium Python, Available from: https://python-visualization.github.io/folium/ [Accessed: 2021-06-01]

[17] Kirkpatrick DG, Seidel, R. The ultimate planar convex hull algorithm?. SIAM journal on computing. 1986;15.1: 287-299.

[18] Fortune S. A sweepline algorithm for Voronoi diagrams. Algorithmica. 1987; 2.1:153-174.

[19] 2011 Administrative Manual, Ministry of the Interior and Safety, Available from: https://www.mois.go.kr/frt/bbs/type001/commonSelectBoardArticle.do?bbsId=BBSMSTR_000000000055&nttId=77460 [Accessed: 2021-06-01]