**Assignment 8(DS-670)- Capstone**

**Midterm-Proposal**

Neha Kumari

**Contribution:**

The analysis is based on the air pollution data from the City Pulse project. There were five air pollutants – Ozone, Carbon Monoxide, particulate matter, sulfur dioxide and Nitrogen Dioxide. Apart from these, there are timestamp, longitude and latitude in the data set.

*Data Preparation*:

As a first step, some data preparation was needed. Each file in the data set was the measure of these five air pollutants at a specific location but during different hours of the day. There are total 221 files; so to sum up the arithmetic mean values of these pollutants were taken. These values represented the average air pollutant levels at a certain location (longitude and latitude).

The data set has been collected from the city plus project which can be found on the following website: <http://iot.ee.surrey.ac.uk:8080/datasets.html>

*Exploratory Data Analysis:*

Scatter plots between the pollutants to understand the relationships between them. Box plots; histogram, time series, scatters plots as relevant for the study. Studying the descriptive statistics and analysis of the data.

* Ozone: Exposure to high concentration of ozone may cause respiratory problems both in adults and children. The symptoms coughing, nausea and chest pain are also common due to high concentration of ozone in the air.
* Carbon Monoxide: It enters into the blood stream and reduces the supply of oxygen to the body organs and tissues, which has adverse effects to human body and mind specially infants.
* Sulfur Dioxide: The major health problems associated with high concentrations of sulfur dioxide include respiratory and breathing problems, cardiovascular disease. People who are asthmatics are more sensitive and prone to its effect.
* Nitrogen Dioxide: It has effects on the ecosystems and on human health as well. It is mainly responsible for respiratory diseases both in adults and in children.
* Particulate Matter: Lung disorders, influenza, respiratory disorders and asthma are the problems associated with particulate matter on the human health.
* As part of data preparation for each attribute (the air pollutant), arithmetic mean is done for that particular location. This means for each file (which represents a specific area in the city) the arithmetic mean of each air pollutant is done to represent the average value of the air pollutant level in that region.

*Unsupervised learning:*

* The K-means algorithm takes input data set and based on the number of clusters, k given as input (eg 3,4,5,6 etc) groups the data, which is found to be similar.
* The objects in the data set is assigned to the cluster it is most closest (closeness is measured as the distance between the object and the centroid, that is the mean of all objects in the cluster).
* The clustering of data is done in such way that the intra cluster similarity is high whereas the inter cluster similarity is low.
* The clustering continues until the centroid of each cluster becomes stable. This means the centroid of the cluster does not change considerably upon introducing a new point in the cluster.
* By applying K-means clustering on the aggregated data we would try to locate the safest area in the city based on the air pollutant levels, similarly locating the unsafe area in the city.
* Since K is an input parameter, we will try with different values of K that gives us more meaningful and clear results. Usually the value of K should neither be very large nor too small.

The data produced by IoT (internet of things) is enormous and data mining techniques can be used to get hidden information, which is of high business value. Smart cities are completely based on IoT. Air pollution is increasing rapidly in the smart cities and has adverse effects on human health. The sources of pollution are many including road traffic, industrial gases and others. In this study we try to find the healthiest areas, which are suitable for leaving, in the smart cities by using K-means clustering. The dataset is generated from the City Plus project. The data is enormous and dynamic due to the number of sensors deployed in the same location and their measurement frequency.

**State of art**:

Many data scientists have explored air pollution data and some of the work done before is as follows:

* The most common type of analysis done so far has been the exploratory analysis. The summary statistics like the mean of each air pollutant, time series analysis to see if the pollution level shows a pattern as per the seasons. The study so far done shows that during monsoons the level of pollution is less in general. Histograms and box-plots have been used as powerful tools for visual analysis. In our analysis the dataset is not suited for such time series analysis as it is based on different hours of the day and not yearly based. However it would be interesting to explore the levels pollution during different hours of the day. This so far has not been included in our study.
* The other type of work done has been done so far is predictive model-supervised learning. The most common models used are multi linear regression model and neural networks. The objective of predictive modeling is to predict the next day pollution levels based on the previous levels. Here the dependent variables are the air pollutants and the independent variables are the environmental factors like temperature, wind speed, humidity etc. Since in our data set the environmental factors that may contribute to the pollution levels are not there so creating out such analysis out of scope. Forecasting is other useful analysis that can tell the future air quality of the area. One thing to note here is that for regression analysis necessary data transformation must be done (like box-cox transformation or any other as relevant). Also we need to check that the conditions for regression are met. The data is normally distributed, error has constant variance and the errors are normally distributed. If the conditions are not met then the model may not give meaningful results. Also graphical analysis of the data is of immense importance in regression and not just descriptive. The model may give strong fit but it may happen that the data is actually not fit for the regression analysis. This can be seen using histograms and scatter plots of the residuals. Also qq plots in R can be useful.
* Neural networks (NN) have also been used for predictive modeling. The neural network models would be more stable model compared to the regression models. In applying neural network generally no data transformation is needed.
* Unsupervised – K means clustering: This analysis is useful for grouping similar data to get some meaningful insights. Like in our study we chose this type of study, as it is suited for our data set. K-means clustering can be used to get the healthy areas in the city. Going further this analysis can also be expanded to understand the traffic situations in the areas. As traffic and pollution have a direct connection, this study can be used to investigate that areas with low pollution may have less traffic congestion and vice-versa. City Plus project is also working on the road traffic dataset, so these analysis combined together can help in building a big picture of the smart cities problems and dealing with them using these analysis. To control the situation there can be some notifications provided to the people about the air quality and traffic situations so that people can take some alternative routes and thus help in worsening the situation in that area. Similarly some other analysis can be done using k means clustering depending on the data set.
* PFCM clustering: Another work done is using PFCM (probabilistic fuzzy c means), this is a measure to get a combine mean from the different areas. This helps in setting a threshold contingency level of the area.

**Data:**

The data is collected from the CityPulse Project <http://iot.ee.surrey.ac.uk:8080/datasets.html>.

The data produced by IoT (internet of things) is enormous and data mining techniques can be used to get hidden information, which is of high business value. Smart cities are completely based on IoT. Air pollution is increasing rapidly in the smart cities and has adverse effects on human health. The sources of pollution are many including road traffic, industrial gases and others. In this study we try to find the healthiest areas, which are suitable for leaving, in the smart cities by using K-means clustering. The dataset is generated from the City Plus project. The data is enormous and dynamic due to the number of sensors deployed in the same location and their measurement frequency.

This data consists of 5 air pollutants namely – ozone, sulfur dioxide, nitrogen dioxide, carbon monoxide and particulate matter. There are 3 more fields in the data set namely- Longitude, latitude and timestamp. The main objective of the CityPulse project is to be able to use this real time data for building real time applications.

Smart city data is not only huge in volume; also it is multi model, changes in formats, representation form, quality and levels of dynamics. City Pulse aims to offer large-scale real time solutions to interlink data from IoT and associated networks and to achieve real-time information for the efficient and smart city applications. The

Smart cities are evolving into larger systems that were earlier disconnected. More and more services and applications in these projects are going to be interlinked. Nowadays-huge amounts of valuable data and sensor information still unused or restricted to certain service domains due to the large number of specific technologies and formats (like

Parking spaces, traffic information, bus timetables, waiting times at events, event calendars, environment sensors for pollution or weather warnings, GPS databases). The aggregation of information from various sources is typically done manually and the collective data is often static. CityPulse will speed up the creation and establishment of valid real-time smart city applications by accumulation two or more disciplines of knowledge-based computing and reliability testing.

Road traffic makes a significant provision to the following emissions of pollutants: benzene (C6H6), nitrogen dioxide (NO2), carbon monoxide (CO), lead, Ozone (O3), particulate matter (PM10 and PM2.5) and sulfur dioxide (SO2). The impact of local air pollution on the environment and human health have been studied and well documented.

**Method:**

Air pollutants have a definite impact on human health condition. Various Data Mining algorithms such as clustering are used to analyze the effects of air pollutants. Also these techniques are used to understand the relationship between weather conditions, road traffic and other factors that might have impact on the air pollutant levels.

K-means algorithm/ Unsupervised learning:

The data set consists of 221 files; each file represents the air pollutant levels of a place in a city during various time intervals and days (which means the latitude and the longitudes remain constant in each file). Since the data huge there is some data transformations needed to work with the data and get some meaningful output from the algorithm.

So we need to aggregate each attribute in each file. We do this by taking the arithmetic mean of each attribute. Now this arithmetic mean will represent the overall air pollutant level of each attribute at a certain location in the city.

X = ∑ xi / n

Where,

X = arithmetic mean of an air pollutant in certain location

n = number records in each file for that location

The K-means algorithm is a simple, unsupervised learning algorithm. It takes the input data set D and the input parameter, K. K is the number of clusters we want to group our data in. While any value of K can be chosen for the algorithm to run, but the results may not be useful. Ideally the value of K should neither be too small nor too large. In our study we will try for different values of K to understand the results.

The grouping of the data in K-means clustering depends on the similarity basis. The partition of the data in K clusters is done in such a way that the inter cluster similarity is low but the intra cluster similarity is high. The K-means algorithm works as follows:

First it randomly choses K objects, each of which initially defined as cluster mean or cluster centroid. For the remaining of the objects each object is assigned to these K objects to which it is most close (The closeness is measured in terms of Euclidean distance). It then re-calculates the cluster mean for each cluster also called the centroid. The process is repeated until there is no major change in the mean value of cluster. This phenomenon is called convergence.

The main steps of k-means algorithm are summarized below:

1. Let X = {x1, x2, x3, x4…xn} be the data points in the data set
2. Randomly assign objects in the data point to these K clusters. These data points are the initial centroids of each k clusters.
3. For the remaining objects, calculate the distance between each data point and the cluster centers.
4. Assign the data point to that cluster whose distance between the cluster center and the data point is minimum.
5. Now again calculate the centroid of each cluster using:

1/ki ∑xi

Where,

Ki represents the number of data points in each cluster.

1. Now again calculate the distance between each data point and obtain new data centers.
2. Continue this process until no data point is reassigned to a new cluster.
3. Since K is an input parameter, we will try with different values of K that gives us more meaningful and clear results. Usually the value of K should neither be very large nor too small.

In this study the main objective to use K-means is to find the healthiest area in the city. The clustering is done to find the minimum levels of the air pollutant. Since we are dealing with five air pollutants, the area having minimum concentration of ozone is considered as the healthiest area.

Note that we can use the levels of other air pollutant also (sulfur dioxide, nitrogen dioxide, carbon monoxide and particulate matter, it is just a matter of choice). Ozone is not emitted directly in the environment by some sources. It is created as a result of the reaction of nitrogen oxide with the sunlight and some volatile organic compounds in the environment. The reactions are encouraged by the sunlight and temperature. So the weather and amount of sunlight has a vital role in the concentration levels of ozone. Emissions from vehicles and other stationary sources can form high ozone concentrations, which can spread, over large regions. For these reasons the ozone level is taken as the deciding factor in finding the healthy and unhealthy area in this project.

Advantages:

1. While there are many unsupervised clustering algorithms but K-means algorithm is fast, easy to understand and robust in nature.
2. It is relatively efficient algorithm. Lets suppose k is the number of clusters, n is the number of objects, d is the dimension of each object in the data set and t is the number of iterations, so generally k, and t, d n

Disadvantages:

1. This algorithm needs an apriori specification for the number of cluster means.
2. If there are overlapping data in the data set, k-means fails to resolve the ambiguity. That means it won’t be able to identify as two separate clusters.
3. The algorithm behaves differently with different representations of data. This means if the data is suppose Cartesian co-ordinates and polar co-ordinates, this will give different results.
4. The Euclidean distance measures can be of unequal weights because of various underlying factors.
5. This algorithm provides only local minima of the squared error function and not the global minima. This is because the initial data points, k is randomly chosen.
6. The algorithm can only be used for continuous data that means it doesn’t work for categorical data.
7. This algorithm also is not fit for data consisting of outliers and noise. This means data cleaning is necessary to make the best out of this algorithm. In our study, the data is structured, cleaned with no missing values.
8. The algorithm works for linear data set. It is unable to handle non-linear data set.

In our study, the data set meets most conditions for the K-means to work well. The data is clean, structured and complete.

We are using Zeppelin platform for this implementation. So far the arithmetic mean of each of the five-air pollutant is taken in each file. This will represent the pollution level at a certain location. Next the task was to combine this aggregated data into a single frame.

To begin with, there has been exploratory data analysis done. Starting with the box plots. The box plots gave the insights of each air pollutant. It helps in understanding the median value and the presence of outliers in the data set if any. The initial study has revealed that the particulate matter and sulfur dioxide to be highly skewed. Also there are some outliers in the data set. Mainly in ozone, particulate matter, carbon monoxide and sulfur dioxide. In general the range of ozone and sulfur dioxide is high in the environment compared to other air pollutants.

The histograms have been plotted for each attribute and going further time series and scatter plots would be used as relevant for the study. K-means will be implemented in python and the results will be compared with different values of K. With the initial study we will start with the initial value of K = 3 and may be will go till K= 10, depending on the results. Most probably we should get the most optimum results within these K values.

With k = 10 we are getting the best results, that is cluster with minimum ozone level concentration. This area can be regarded as the healthiest area in the smart city.

In this study the main objective to use K-means is to find the healthiest area in the city. The clustering is done to find the minimum levels of the air pollutant. Since we are dealing with five air pollutants, the area having minimum concentration of ozone is considered as the healthiest area.

Note that we can use the levels of other air pollutant also (sulfur dioxide, nitrogen dioxide, carbon monoxide and particulate matter, it is just a matter of choice). Ozone is not emitted directly in the environment by some sources. It is created as a result of the reaction of nitrogen oxide with the sunlight and some volatile organic compounds (VOC) in the environment. The reactions are encouraged by the sunlight and temperature. So the weather and amount of sunlight has a vital role in the concentration levels of ozone. Emissions from vehicles and other stationary sources can form high ozone concentrations, which can spread, over large regions. For these reasons the ozone level is taken as the deciding factor in finding the healthy and unhealthy area in this project.

Road traffic makes a significant provision to the following emissions of pollutants: benzene (C6H6), nitrogen dioxide (NO2), carbon monoxide (CO), lead, Ozone (O3), particulate matter (PM10 and PM2.5) and sulfur dioxide (SO2). The impact of local air pollution on the environment and human health have been studied and well documented

**References:**

* “CityPulse project” and the URL is http://iot.ee.surrey.ac.uk:8080/datasets.html
* Ganti, R. K., Ye, F., & Lei, H. (2011). Mobile crowdsensing: current state and future challenges. *IEEE Communications Magazine*, *49*(11).
* Lu, W., Wang, W., Leung, A. Y. T., Lo, S. M., Yuen, R. K., Xu, Z., & Fan, H. (2002). Air pollutant parameter forecasting using support vector machines. In *Neural Networks, 2002. IJCNN'02. Proceedings of the 2002 International Joint Conference on* (Vol. 1, pp. 630-635). IEEE.
* Wang, W., & Guo, Y. (2009, October). Air pollution PM2. 5 data analysis in Los Angeles long beach with seasonal ARIMA model. In *Energy and Environment Technology, 2009. ICEET'09. International Conference on* (Vol. 3, pp. 7-10). IEEE.
* Seaton, A., Godden, D., MacNee, W., & Donaldson, K. (1995). Particulate air pollution and acute health effects. *The lancet*, *345*(8943), 176-178.
* Sowe, S. K., Kimata, T., Dong, M., & Zettsu, K. (2014, July). Managing heterogeneous sensor data on a big data platform: IoT services for data-intensive science. In *Computer Software and Applications Conference Workshops (COMPSACW), 2014 IEEE 38th International* (pp. 295-300). IEEE.
* Yao, L., Lu, N., & Jiang, S. (2012, May). Artificial neural network (ANN) for multi-source PM2. 5 Estimation using surface, MODIS, and meteorological data. In *Biomedical Engineering and Biotechnology (iCBEB), 2012 International Conference on* (pp. 1228-1231). IEEE.
* Osborne, M. A., Roberts, S. J., Rogers, A., & Jennings, N. R. (2012). Real-time information processing of environmental sensor network data using bayesian gaussian processes. *ACM Transactions on Sensor Networks (TOSN)*, *9*(1), 1.
* Maitra, R. (2009). Initializing partition-optimization algorithms. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, *6*(1), 144-157.
* Gionis, A., Mannila, H., & Tsaparas, P. (2007). Clustering aggregation. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, *1*(1), 4.
* Kanungo, T., Mount, D. M., Netanyahu, N. S., Piatko, C. D., Silverman, R., & Wu, A. Y. (2002). An efficient k-means clustering algorithm: Analysis and implementation. *IEEE transactions on pattern analysis and machine intelligence*, *24*(7), 881-892.
* Pal, N. R., Pal, K., Keller, J. M., & Bezdek, J. C. (2005). A possibilistic fuzzy c-means clustering algorithm. *IEEE Transactions on Fuzzy Systems*, *13*(4), 517-530.
* Jerrett, M., Burnett, R. T., Ma, R., Pope III, C. A., Krewski, D., Newbold, K. B., ... & Thun, M. J. (2005). Spatial analysis of air pollution and mortality in Los Angeles. *Epidemiology*, *16*(6), 727-736.
* Darken, C., & Moody, J. (1990, June). Fast adaptive k-means clustering: some empirical results. In *Neural Networks, 1990., 1990 IJCNN International Joint Conference on* (pp. 233-238). IEEE.
* Huang, Z., & Ng, M. K. (1999). A fuzzy k-modes algorithm for clustering categorical data. *IEEE Transactions on Fuzzy Systems*, *7*(4), 446-452.
* Zhang, Z., Zhang, J., & Xue, H. (2008, May). Improved K-means clustering algorithm. In *Image and Signal Processing, 2008. CISP'08. Congress on* (Vol. 5, pp. 169-172). IEEE.
* Graves, T. L., Harrold, M. J., Kim, J. M., Porter, A., & Rothermel, G. (2001). An empirical study of regression test selection techniques. *ACM Transactions on Software Engineering and Methodology (TOSEM)*, *10*(2), 184-208.
* Kampa, M., & Castanas, E. (2008). Human health effects of air pollution. *Environmental pollution*, *151*(2), 362-367.
* Selden, T. M., & Song, D. (1994). Environmental quality and development: is there a Kuznets curve for air pollution emissions?. *Journal of Environmental Economics and management*, *27*(2), 147-162.
* Zivkovic, Z., & van der Heijden, F. (2004). Recursive unsupervised learning of finite mixture models. *IEEE Transactions on pattern analysis and machine intelligence*, *26*(5), 651-656.
* Le, Q. V. (2013, May). Building high-level features using large scale unsupervised learning. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on* (pp. 8595-8598). IEEE.