**Assignment – 5(DS-670)**

**Algorithm Overview and Implementation**

**CITY PLUS Pollution Dataset**

**Neha Kumari**

**INTRODUCTION**:

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.

The impact of air pollution on health has been well studied and is summarized below:

Social & Economic Impact

Road Traffic

Exposure Level

Health Problems

Dose Response

Bad Air Quality

Humans

Air Pollutants

**METHODOLOGY**:

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 learing:

The data set consists of 449 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

Algorithm Description:

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 (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.

Advantages of using K-means:

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 of using K-means:

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.

Analysis so far:

We are using Zeppelin platform for this implementation. So far the arithmetic mean of each of the 5 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.

**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.