ASSGNMENT -3 (DS-670)

Expected Contribution and State of art

CITY PLUS Pollution Dataset

Neha Kumari

**Introduction:**

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.

**Expected Contribution**:

The analysis is based on the air pollution data collected from the city plus project. There are five air pollutants in the data set – Ozone, Carbon Monoxide, particulate matter, Sulphur Dioxide and Nitrogen Dioxide. In addition there is longitude, latitude and timestamp field. There are total 449 files in the data set; each file has the measure of these five air pollutants at a specific location but during different hours of the day. The expected analysis on the data set is :

1. **Data description and preparation**:

The data set consists of five air pollutants, latitude, longitude and a timestamp field. The various air pollutants and it’s effect are explained below:

* 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.
* Sulphur Dioxide: The major health problems associated with high concentrations of sulphur dioxide include respiratory and breathing problems, cardiovascular disease. People who are asthmatics are more sensitive and prone to it’s 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.
* Data preparation: 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>
* The dataset has total 449 files. Each file represents the air pollutant level of a particular area in the city at different time of the day. This means in each file the latitude and longitudes are fixed for each file.
* 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.

1. **Exploratory data analysis**:

* To find if there exists Correlation between the various air pollutants using correlation matrix.
* Box plots; histogram, time series, scatters plots and other plots as relevant for the study.
* Descriptive statistical analysis: chi square test for independence, hypothesis tests to compare the mean values of the air-pollutants, t-tests as relevant for the study.

1. **Unsupervised Learning- K-means clustering**:

* 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.

**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.

**References:**

* “CityPulse project” and the URL is http://iot.ee.surrey.ac.uk:8080/datasets.html
* B. Ojeda-Magaña1, M. G. Cortina-Januchs2, J. M. Barrón-Adame, J. Quintanilla-Domínguez, W. Hernandez,A. Vega-Corona, R. Ruelas1 and D. Andina “Air pollution Analysis with a PFCM Clustering Algorithm Applied in a Real Database of Salamanca (Mexico)” 978-1-4244-5697- 0/10/$25.00 ©2010 IEEE
* Jes Fenger: “Air pollution in the last 50 years - From local to global”.Journal of Atmospheric Environment, vol. 43, pp. 13-22, 2009
* J. M. Barron-Adame, J. A. Herrera Delgado, M.G. Cortina-Januchs,D. Andina, A. Vega-Corona: “Air Pollutant Level Estimation Applyinga Self-organizing Neural Network”.Proceedings of the 2nd interna-tional work-conference on Nature Inspired Problem-Solving Methods inKnowledge Engineering. IWINAC ’07, pp. 599-607, 2007
* G. Cervone, P. Franzese, Y. Ezber, Z. Boybeyi, F. Bonchi, B. Berendt, F. Giannotti, D. Gunopulos, F. Turini, C. Zaniolo, N. Ramakrishnan and X. Wu, “Risk Assessment of Atmospheric Hazard Releases using K-means Clustering”, Proceedings of IEEE International Conference on In Data Mining Workshops, (2008) December 15-19; Pisa, Italy.
* Samet, J.M., Zeger, S.L., Kelsall, J.E., Xu, J. and Kalkstein, L.S. (1997), Air Pollution, Weather and Mortality in Philadelphia, 1973{1988. In Particulate Air Pol lution and Daily Mortality: Analyses of the E  
  ects of Weather and Multiple Air Pol lutants. The Phase IB Report of the Particle Epidemiology Evaluation Project. Health E  
  ects Institute, Cambridge MA, pp. 1{29.
* Yi J. and Prybutok V. R. (1996), A neural network model forecasting for prediction of daily maximum ozone concentration in an industrialized urban area, Environmental Pollution, 92, 349-357.
* B. Ojeda-Magaña, R. Ruelas, F.S. Buendía-Buendía and D.Andina. “AGreater Knowledge Extraction Coded as Fuzzy Rules and Based on theFuzzy and Typicality Degrees of the GKPFCM Clustering Algorithm”.In Intelligent Automation and Soft Computing, Vol. 15(4): 555–571,2009
* J. M. Barron-Adame, J. A. Herrera Delgado, M.G. Cortina-Januchs,D. Andina, A. Vega-Corona: “Air Pollutant Level Estimation Applyinga Self-organizing Neural Network”.Proceedings of the 2nd interna-tional work-conference on Nature Inspired Problem-Solving Methods inKnowledge Engineering. IWINAC ’07, pp. 599-607, 2007
* A. Mace, R. Sommariva, Z. Fleming and W. Wang, in Adaptive K-means for clustering air mass trajectories, Edited H. Yin, W. Wang and V. Rayward-Smith, Springer Berlin, Heidelberg, vol. 6936, (2011), pp. 1-8.
* L. Li and S. Cheng, “A Calculated Methodology of Regional Contributions Based on MM5-CAMx in Typical City: A 2006 Case Study of SO2 and Sulfate”, Proceedings of the 4th International Conference on Bioinformatics and Biomedical Engineering, (2010) June 18-20; Chengdu, China