Project Report

INNOPOLIS UNIVERSITY

Air Clustering

Student	Klim Basargin
Course	Big Data Technologies and Analytics
Semester	Spring
Year	2023

Contents

1. Introduction

Environmental issues are very relevant today. The main reason is all kinds of production and factories. Their emissions pollute the atmosphere and spoil natural conditions. To solve this problem, various methods are used, some of them work, some do not. Nevertheless, much is being done today to save the environment.

2. Business Understanding

In order to punish and motivate companies, it was decided to test a dynamic taxation system. The amount of tax payments will correspond to the quality of emissions generated by enterprises. As a data scientist, it is required to accurately define air quality classes so that when assessed in the future, one can unequivocally say what tax rate to give a particular factory.

2.1 Current situation assessment

Project now is only about Beijing air quality. For other cities I can not guarantee scalability.

2.1.1 Inventory of resources

List the resources available to the project including:

- Personnel (one data scientist student)
- Data (fixed extracts, access to live, warehoused, or operational data)
- Computing resources (hardware platforms)
- Software (data mining tools, other relevant software)

2.1.2 Requirements, assumptions and constraints

We need to get clusters of air quality. We do not know how many them should be. Clusters centers should be saved.

2.1.3 Risks and contingencies

There is a risk that the data cannot be clustered, or the number of clusters will be too large, more than 10.

2.1.4 Costs and benefits

This will improve the environment in Beijing and correct the fairness of taxation.

2.1.5 Business Objectives

Clearly separate types of emissions from industries and apply the correct tax rates.

2.2 Data mining objectives

Achieve strong clustering.

2.2.1 Business success criteria

Retrieving multiple explicit groups of air quality.

2.2.2 Data mining success criteria

Silhouette criterion greater than 0.75.

2.3 Project plan

1. Upload dataset
2. Explore data
3. Transform data
4. Modeling and tuning
5. Evaluation

3. Data Understanding

To complete the task, air data for a long period are required, it is necessary to evaluate the content of various substances, physical and chemical properties.

3.1 Initial data collection

Data set includes hourly air pollutants data from 12 nationally-controlled air-quality monitoring sites. The air-quality data are from the Beijing Municipal Environmental Monitoring Center. The meteorological data in each air-quality site are matched with the nearest weather station from the China Meteorological Administration. The time period is from March 1st, 2013 to February 28th, 2017.

3.1.1 Big data pipeline: Stage I

I imported data from .csv file to PostgreSQL table. "Stage1.sh" script does these tasks.

3.2 Data Description

Features:

No: row number

year: year of data in this row

month: month of data in this row

day: day of data in this row

hour: hour of data in this row

PM2.5: PM2.5 concentration (ug/m$\hat{3}$)

PM10: PM10 concentration (ug/m$\hat{3}$)

SO2: SO2 concentration (ug/m$\hat{3}$)

NO2: NO2 concentration (ug/m$\hat{3}$)

CO: CO concentration (ug/m$\hat{3}$)

O3: O3 concentration (ug/m$\hat{3}$)

TEMP: temperature (degree Celsius)

PRES: pressure (hPa)

DEWP: dew point temperature (degree Celsius)

RAIN: precipitation (mm)

wd: wind direction

WSPM: wind speed (m/s)

station: name of the air-quality monitoring site

Data description report – Dataset has 420768 records and 18 features. It contains time-series data. Also it has 53728 N/A values.

Big data pipeline: Stage II The creation of Hive table is in "Stage2.sh".

3.3 Data exploration

I explored balance of data between stations, it is equal. Also I looked at wind direction frequency in years and mean of temperature on each station.

Data exploration report – I have 35064 rows from each station. Air characteristics holds every year and are similar between stations.

Big data pipeline: Stage II All connected are able in outputs folder.

3.4 Data quality

The data is pretty clean, no outliers, no imbalance.

4. Data Preparation

This is the stage of the project where you decide on the data that you're going to use for analysis. The criteria you might use to make this decision include the relevance of the data to your data mining goals, the quality of the data, and also technical constraints such as limits on data volume or data types.

4.1 Data selection

In clustering I need all data connected with quality, time to order rows and station to separate records.

Rationale for inclusion/exclusion – No features for exclusion.

4.2 Data cleaning

I imputed all real numbers by mean and strings by most frequent value. I encoded wind direction to numbers.

Data cleaning report – 53728 cells were imputed.

4.3 Data construction

I used PCA to reduce all features except station and datetime columns to one in order to get one number that will explain timestamp. Then I create new dataset from slides. Every slide is 24 serial rows from one station. After that I get dataset which consists of vectors that characterize the period of time. Now I do not need station name or timestamps. From this step I can try to clusterize time-series data.

5. Modeling

5.1 Select modeling technique

I preferred models that work on the basis of centroids, because later, with new data, you can very quickly determine which cluster it belongs to. I need only to save coordinates of centroids.

Modeling technique – K-Means and Bisecting K-Means.

5.2 Build model

I used pyspark.ml to implement and evaluate performance.

Parameter settings – I change number of clusters and seed.

5.3 Assess model

Both models show the similar results.

Model assessment – 2 clusters was chosen as the best number.

Revised parameter settings – I ran models from 2 to 5 clusters and with different seeds.

Big data pipeline: Stage III All modeling part is in ”pipline.py” file and information about coordinates of clusters saved to model folder.

6. Evaluation

I evaluate models with Silhouette coefficient. It changes betweet 1 and -1. Where 1 is the best clustering.

Approved models – I got Silhouette coefficient equals to 0.77 with K-Means and 0.8 with Bisecting K-Means.

Big data pipeline: Stage III All this plots are in outputs folder.

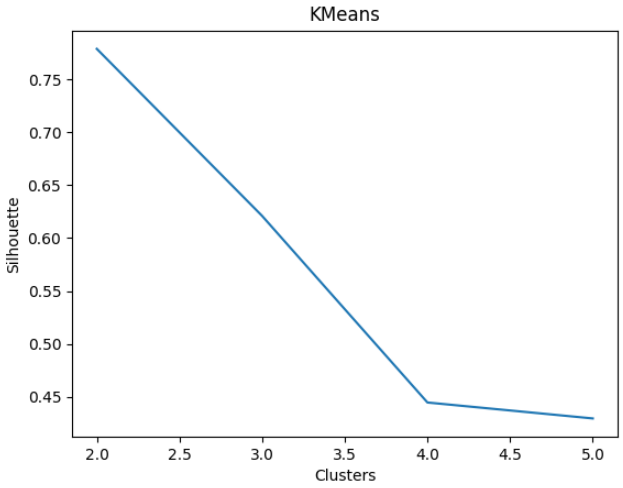

Figure 1: Evaluation of K-Means

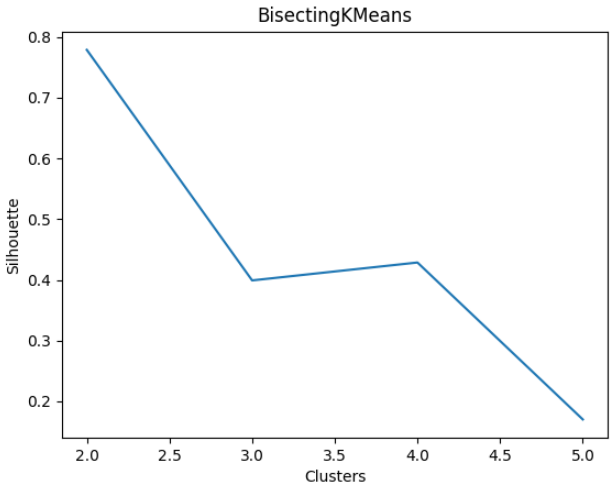

Figure 2: Evaluation of Bisecting K-Means

7. Deployment

Of course, I cannot decide what kind of air is bad or not, requires expert work to determine how to characterize the air from each cluster. But now all need is to transform new data and look to what cluster it belongs to.

Deployment plan –

1. Explore clusters with experts in environmental problems.
2. Determine what conditions will be imposed on enterprises.
3. Analyze new real life examples.

7.1 Limitations and Challenges

I were limited by only one big city. In future it is important to analyze more samples from all world.

8. Contributions and Reflections on own work

8.1 Report summary

- I explored air quality data.
- Tried different approaches to reach results.
- I improved understanding of time-series data analyzing.
- In future it is important to analyze more samples.

References

Alqahtani, Ali, et al. ”Deep time-series clustering: A review.” Electronics 10.23 (2021): 3001.