Data Mining (CS F415) Project Final Report

Group Number: 12

24th November 2017

Project Title: Kaggle: World Development Indicators

Team Size: 4 **Group Number: G12**

Team Members:

Anant Sharma 2014A7PS0051G, Snimarpal Singh 2014B3A70646G,
Gyanesh Malhotra 2014B3A70495G, Ranajoy Roy 2014B4A70604G

Objective

To infer non-intuitive relations between the various attributes and accordingly cluster the records in the given data set after trimming, exploring, preprocessing and transforming the data appropriately.

To cluster countries on the basis of these attributes and to perform intercluster and intracluster analysis to understand the contrast between the clusters and the similarities within a cluster respectively.

Initial Approach:

We downloaded the proposed data set, World Development Indicators, from Kaggle. It was mentioned to be the most accurate global development data available. Kindly visit:
https://www.kaggle.com/theworldbank/world-development-indicators

The data itself amounted to about 234 MB after the CSVs were extracted from the zipped folder. We studied the various CSV files available:

1. WDICountry-Series.csv : This file lists the data sources for the various countries which are a part of the WDIData

2. WDICountry.csv : This serves as a country code lookup, it mentions the countries along with their codes and a few basic geographic and survey related details

3. WDIData.csv : This is the time series data which lists out the values of various world development indicators (attributes) for countries (records) over the years (1960-2016)

4. WDIFootNote.csv, WDISeries-Time.csv, WDISeries.csv : These include footnotes, methodology notes and, series data definitions & specific metric details respectively

Phases of the Framework worked upon for Demo I:

1. Data Collection and Trimming the Data set:

    - We played around with the data to estimate which attributes are essential for the latter steps in the project and which time period should be considered.

2. Exploratory Data Analysis:

    - We obtained graphs that helped us better visualize the data and determine how we must proceed further.
    - We defined clear metrics for the data, checked for outliers in the data set and ascertained whether or not they should have been considered while creating the model.

Learning Outcomes of the Demo I Tasks

- After going through the various CSVs, we decided to compute the sparsity of the data. This was done on Jupyter notebook while keeping the time period and the countries in mind.

- Finally, the time period 1975-2015 was chosen for further analysis and 7 countries were picked for EDA. This was done on the basis of their development status and the sparsity of their data (for both the entire period and 1975-2015 [double verification]).

- Later, with the help of interpolation and scatter plots on matplotlib a comprehensive representation of each attribute was possible which helped appreciate the difference between the various countries chosen.

- 24 attributes spanning across various development related fields (Finance, Education, Import-Export, Environment, Manufacturing, etc.) were chosen to get a wholesome idea of the country's developmental status

Status after Demo I:

By the time of the first milestone evaluation in September we had successfully chosen our data set and performed an elaborate Exploratory Data Analysis to better understand the data set.

Since the data set was huge (more than 200 attributes for 150+ countries across the years 1960-2015) and sparse at the same time (with more than 50% of the values missing for select attributes of some countries) it was imperative that we chose the sample set wisely for further analysis.

Phases of the Framework worked upon for Demo II:

1. Preprocessing of the data:
    - By choosing the countries based on the sparsity index from last time, the features and years based on the availability of data, the need for cumbersome preprocessing was obviated with the exception of the interpolation of a few missing with the help of the python library pandas
2. Clustering of Countries for each chosen year and feature bucket:
    - K-means was used since the data set for each clustering iteration was relatively small (therefore, Density based clusters wouldn't make sense)

Narrowing down the Data before Clustering

1. Country Selection: We chose 3 groups of developed, developing and underdeveloped countries, i.e. G7, BRICS, and ASEAN respectively to increase the diversity of the sample data set.
2. Feature Buckets : The World Bank data had a spreadsheet categorizing the attributes into different classes such as (Economic, Infrastructure, Public and Private Sector, Environment, Health, Education, Poverty, etc.). We considered these buckets for our clustering analysis.

3. Attribute Selection : Once the feature buckets were decided we selected attributes in each bucket based on the availability of the data for the countries chosen.

4. Year Selection : Initially we thought of choosing '95, '05 and '15, but then we finalized an event based approach. This would let us focus on years in which the world was affected by occurrences such as the Asian Financial Crisis and the recession in 2008.
   We were able to cross verify the impact of the events and choose the years by plotting the time series graph of the GDP growth rate (annual percentage) for three categories (underdeveloped, developing and developed). The years **1998, 2005 and 2009** were finalized for the cluster analysis.

Learning Outcomes of the Demo II Tasks

- We took K=5, since K=4 gave rise to uneven clusters which had countries that weren't too similar to each other.
- The representative for each cluster was chosen to be the country that was closest to the centroid of the cluster.
- The clusters were plotted in 2 and 3 dimensions after performing PCA on the attributes of the feature buckets. The first two and three principal components were chosen for the plots respectively.
- Various domain specific inferences were drawn from the clusters formed and the comparative analysis of the representative countries of each cluster.

Learning Outcomes of the Final Demo Tasks

1. Cluster Quality Analysis:
   - The two methods applied to test the clusters formed are:
     1. Elbow Method and 2. Silhouette Analysis
   - The **Elbow Method** looks at the percentage of variance explained as a function of the number of clusters. The point at which marginal gain in explained variance will drop is known as the elbow. The number of clusters is chosen corresponding to this point.

- In k-means one must minimize the within cluster sum of squares (WCSS), basically one must minimize:

$$arg_S min \Sigma_{i=1}^{k} \Sigma_{x \in S_i} ||x - \mu_i||^2$$

- The **Silhouette Value** is a measure of how similar an object is to its own cluster (cohesion) compared to other clusters (separation). The silhouette ranges from -1 to 1.
- The closer it is to 1, the more cohesive the cluster. The Silhouette value is given by:

$$s(i) = \frac{b(i) - a(i)}{max\{b(i), a(i)\}}$$

Conclusion

Over the course of the project we were able to implement various techniques that we had learnt in the data mining course (EDA, Clustering, Elbow Analysis, etc.) and we also came across new methods (Sparsity Analysis, Silhouette Analysis, etc.) which we applied to further understand and analyze the chosen dataset.

Some key insights that we obtained after mining the data set include:

(a) The Asian Financial Crisis of 1997 caused drastic changes in the economic attributes of Asian Countries. Indonesia and South Korea had a negative GDP Growth rate in the years that followed and it took them a while to recover from that. (refer Additional Analysis '98)

(b) Another set of attributes worth taking note of is the Health Bucket of African Countries. Nigeria has the highest fertility rate and mortality rate, this keeps the population in check but the age dependency ratio is the highest as well (Therefore, senior citizens need to earn a living) (refer '09 analysis).

(c) Finally, the global recession of '08 caused several developed countries in the west to have a negative GDP Growth rate and this in turn affected other countries as well as the GDP per captita stagnated over the next few years.