Milestone 1 – Unsupervised Learning

Maciej Nalepa
maciej.nalepa@alu.uclm.es

Piotr Maliszewski
piotr.maliszewski@alu.uclm.es

Gökay Iseri
gokay.iseri@alu.uclm.es

Başar Milli
basar.milli@alu.uclm.es

ABSTRACT

The purpose of this work is to explore the environmental data collected by various U.S. Federal Government Agencies from two cities (San Juan, Puerto Rico and Iquitos, Peru) to gain a better understanding of the Denge Spread Phenomena. These data are from a competition of the site DrivenData. Training data will be used. The overall objective is to use unsupervised learning techniques to make a preliminary exploration of the data and to extract conclusions from discarded elements, etc. The specific objectives are as follows:

(1) Identification of outliers elements (weeks) in the dataset.
(2) Use clustering algorithms to identify groups and characterize them.
(3) (optional) Feature Selection using clustering algorithms.

CCS CONCEPTS

• **Computing methodologies → Machine learning**.

KEYWORDS

machine learning, unsupervised

ACM Reference Format:
Maciej Nalepa, Piotr Maliszewski, Gökay Iseri, and Başar Milli. 2020. Milestone 1 – Unsupervised Learning. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

In this project we explore Unsupervised Learning methods. We were working on DrivenData DengAI: Predicting Disease Spread competition dataset. Exactly our group was focused on finding relations among the data from San Juan between years 1992 - 1998.

The code and source files are available at https://github.com/GummyBearStudioTeam/ESI_MLTechniques.

1.1 Correlation factor

The data is not strongly correlated except the temperature fields.

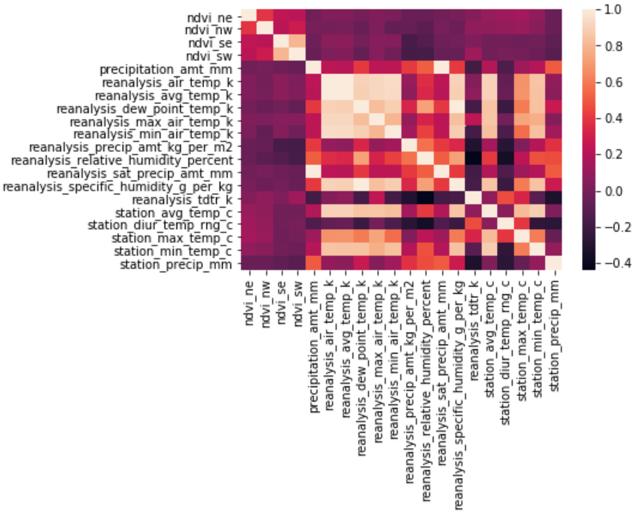

Figure 1: Unprocessed data correlation matrix

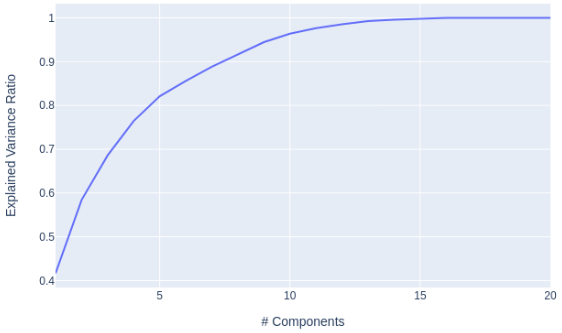

Figure 2: PCA: Explained variance ratio

Table 1: DBSCAN results with specific epsilon.

epsilon	clusters	outliers
1.2	0	364
1.4	1	360
1.6	4	346
1.8	8	297
2.0	5	252
2.2	5	180
2.4	3	134
2.6	2	91
2.8	2	58
3.0	1	37
3.2	1	29
3.4	1	17
3.6	1	10
3.8	1	6
4.0	1	5
4.2	1	3
4.4	1	3
4.6	1	2
4.8	1	2

2 DIMENSIONALITY REDUCTION

According to the explained variance ratio by projecting the data to 10 dimensions, we can preserve arround 96% of information. That allows us to keep a lot of information by reducing total number of dimensions by half.

3 OUTLIER IDENTIFICATION

Features that have the biggest impact on outliers are all of the measures of precipitation(Highest purple, green and blue bars). Differences are noticeable so we decided not to take them into consideration in further analysis.

4 CLUSTERING

The number of cluster (k) has been chosen as 3, because higher values reduce Silhouette score significantly.

4.1 K-means

(1) Group 0 has greater mean of precipitation while having lower temperatures. Humidity is around the average. This group could be labelled as having low temperature with big precipitations

(2) Group 1 has an average precipitation, but it's climate is more humid and temperatures are higher than a mean of the whole data. This group could be labelled as as having high temperature with big precipitations

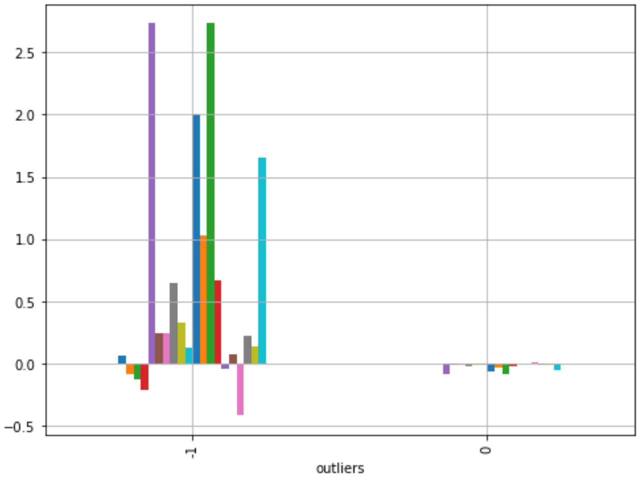

Figure 3: Outliers' mean

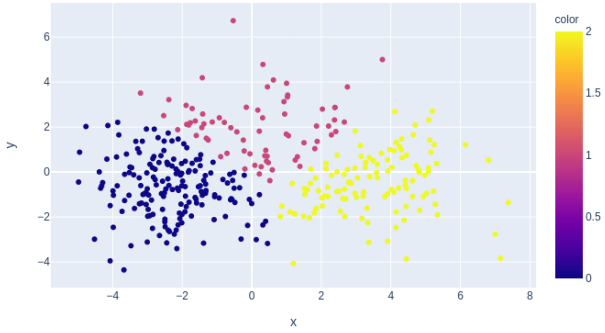

Figure 4: K-means clusters projected into 2D

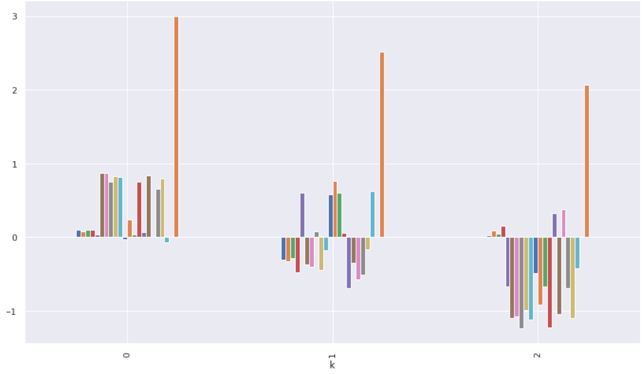

Figure 5: K-means - mean diagram

Conference'17, July 2017, Washington, DC, USA
© 2020 Association for Computing Machinery.
ACM ISBN 978-x-xxxx-xxxx-x/YY/MM...$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

(3) Group 2 has noticeable the biggest humidity from the previous groups. Both temperature and precipitation are low. Almost all features deviate from the mean significantly. This group could be labelled as having low temperature with small precipitations.

4.2 Hierarchical

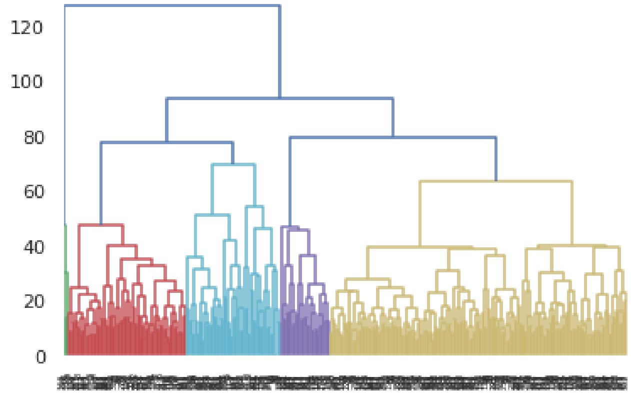

Figure 6: Dendogram

The clusters have been created by cutting a tree at the height 70. It was a value that allowed us to make some bigger insight about the data. Different value either could group everything in gigantic groups or would be too fragmented.

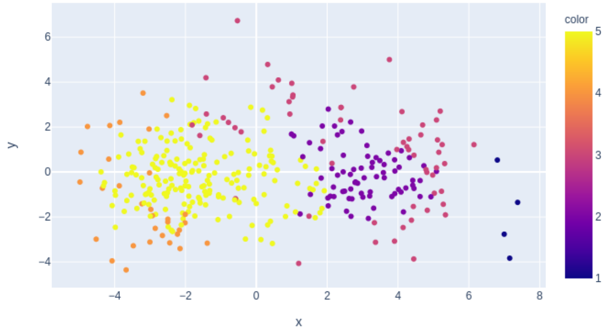

Figure 7: Hierarchical clustering PCA chart

(1) Group 0 is characterised by a greater diurnal temperature range. Winds are above the average when temperatures precipitation and humidity are very low but the whole group is small, so it's mean should not be used to extract further details from it.
(2) Groups 1 and 2 are quite similar. The biggest difference is that the 1st group has lower precipitation when the 2nd one has a precipitation around the global average. Both have low temperatures and humidity.
(3) Group 3 - In that group all of the features are above the average. It seems that the temperature is the most significant factor. It is the only representative group (without taking into consideration the 0th group) where strong winds occurs.
(4) Group 4 - In that group most of features are above the average, but a deviation is smaller than in the 3rd group. What characterise that group as well are winds which force is about the average

5 FEATURE SELECTION

Basing on hierarchical clustered data we propose following features:

- `ndvi_ne`
- `ndvi_nw`
- `ndvi_se`
- `ndvi_sw`
- `precipitation_amt_mm`
- `reanalysis_precip_amt_kg_per_m2`
- `reanalysis_specific_humidity_g_per_kg`

Mean values of these features in each cluster deviate from each other significantly.

6 REFINEMENT

Columns allowing to identify the week were dropped but if they were preserved it could allow to get some conclusions manually.

Variable names for storing data could be improved, because currently they can be difficult to distinguish. (e.g. `df` and `df2`)

Feature selection was performed manually by comparing charts of mean values between clusters, which should instead be performed using some numerical metrics. There was also no verification if the selected features are sufficient to distinguish the clusters in the same way.

K-means, hierarchical clustering and DBSCAN parameters can be adjusted to create different groups.

7 SUMMARY

We have explored the dataset and learned how to use the unsupervised learning tools.

Scatter plots representing the clusters work the better the less clusters we create. The resulting groups are similar when using either method of clustering, we decided to create more clusters using hierarchical method.

ACKNOWLEDGMENTS

This work is part of UCLM project of the Machine Learning Techniques subject. Group PMG.