Milestone 1 Report - Unsupervised Learning

Machine Learning

Agustín Mora Acosta
Agustin.Mora1@uclm.es
UCLM
Ciudad Real, Spain

Andrés González Díaz
andres.gonzalez8@alu.uclm.es
UCLM
Ciudad Real, Spain

ABSTRACT

Unsupervised learning techniques are used to explore the dataset to extract meaningful knowledge without the need for an established objective or reward.

These techniques are not only useful for the discovery of structures in unlabeled data, but are also very useful for the selection of relevant features and data compression (essential tasks before using other automatic learning techniques) function.

In this approach, we used different unsupervised learning techniques to identify outliers elements in the data, identify groups, characterize them, and reduce the dimensionality of the data.

KEYWORDS

datasets, clustering, feature, unsupervised, PCA, outlier

ACM Reference Format:
Agustín Mora Acosta and Andrés González Díaz. 2020. Milestone 1 Report - Unsupervised Learning: Machine Learning. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

The purpose of this work is to explore and analyze the environmental data collected by various U.S. Federal Government Agencies from two different cities (San Juan in Puerto Rico and Iquitos in Peru) to gain a better understanding of the Dengue Spread Phenomena.

This data comes from a competition of the site DrivenData and, for this work, we only consider data from the city of San Juan between 1990 and 1996.

Our main objective is to apply unsupervised learning techniques to make a preliminary exploration of the data and extract some conclusions from discarded elements.

You can find the repository that contains the notebook used to apply these techniques in the following link:

https://github.com/Ofeucor/Machine-Learning-Techniques

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

2 DIMENSIONALITY REDUCTION

The excessive dimensionality of a dataset brings two relevant problems: computational cost of processing the data and more probability of bad data. In this assignment, concerning methodology, we have used dimensionality reduction to reduce the number of elements of the dataset, which consists of transforming the data into a less dimensions space, preserving some information from the original data.

Figure 1: Correlation among features

Before applying dimensionality reductions techniques, we've extracted the correlation among features, to obtain some conclusions. Based on Fig. 1, the conclusion extracted are the followings:

- The features nvdi_ne, nvdi_nw, nvdi_se, nvdi_sw are slightly correlated between them, but they are not correlated to any other feature. Maybe this features could be reduced in one feature.
- The features from the reanalysis related to temperature, are strongly correlated between them, and also strongly correlated with other features as station temperature features. (station_min_temp_c, station_max_temp_c...) and reanalysis_specific_humidity_g_per_kg.

- The features related to precipitation are highly correlated (precipitation_amt_mm, reanalysis_sat_precip_amt_mm and station_precip_mm).
- The relative humidity percent is inversely correlated with the thermal amplitude (reanalysis_tdtr_k) and the diurn temperature range (station_diur_temp_rng_c).

Despite this conclusions, we didn't select a subset of features, we used the Principal Component Analysis (PCA) to project the data on a set of orthogonal dimensions (components) that are linear combination of the original atributtes, preserving the global variance of the data. After executing PCA, we decided to use 3 components to reduce the dimensionality of the data, keeping almost the 80% of explained variance of the data. We extracted some conclusions from this components:

- The first component (PC-1) is generally related to the temperature and humidity in the data.
- The second component (PC-2) is more related to the precipitation, relative humidity and thermal amplitude, despite being slightly linearly related to almost all of the features.
- The third component (PC-3) is highly related to the vegetation features, but also is related to the precipitation in mm.

In the Fig. 2 we can see the representation of the data in a 3D space, using the components obtained by the Principal Component Analysis.

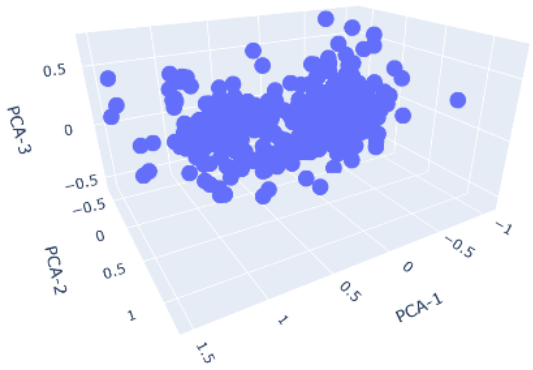

Figure 2: 3D representation of the data

3 OUTLIER IDENTIFICATION

After dimensionality reduction, we explore the data to find outliers. Outliers are observation inconsistent with rest of the dataset, due to excepcional data, poor data quality or low quality measurements. In order to do this outlier search, we used the DBSCAN algorithm.

For the parametrization of this algorithm, the following considerations have been taken:

- Due to the lack of an expert in the domain, we used the ln(n) heuristic approach to set the minPts parameter of the algorithm, where n is the total number of points to be clustered (347 in our case).
- In order to obtain the epsilon parameter for the DBSCAN algorithm, we computed the euclidean distance from each point of the data to its neighbors, and we plotted the sorted distance of every point to its kth nearest neighbor. Then, we tried some epsilons to choose the definitive one, 0.65.

With a parametrization of minPts=6 and eps=0.65, DBSCAN group the data in 1 cluster, detecting 4 outliers.

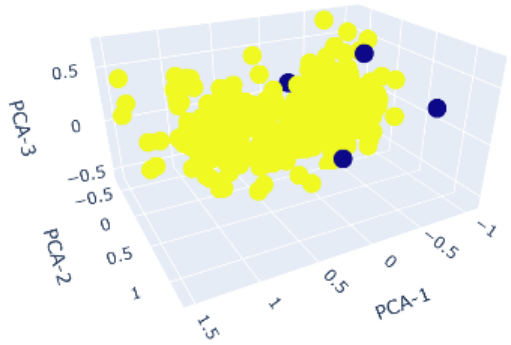

Figure 3: Outlier identification by DBSCAN algorithm

After identifying this outliers, we analyzed why these elements were outliers, in order to decide whether or not consider them for further analysis. These are our conclusions and decisions about this outliers:

- `weekofyear:21 year:1992` - This outlier could be an outlier because of the extreme value in reanalysis_precip_amt_kg_per_m2, which seems like an error in the data because this value does not correspond to the precipitation in mm in the precipitation_amt_mm feature.
- `weekofyear:53 year:1993` - This outlier seems like a week with a great thermal amplitude, but looks factible so well consider this data for further analysis.
- `weekofyear:21 year:1995` - This outlier could be considered outlier because a high value on precipitation_amt_mm and reanalysis_sat_precip_amt_mm but a low value on reanalysis_precip_amt_kg_per_m2. This could be an error in the data so we wont consider this data for further analysis.
- `weekofyear:38 year:1996` - This data looks like an extremely rainy week, so we wont consider this data because it can distort data while aplying another clustering algorithm like Hierarchical Clustering Algorithm.

4 CLUSTERING BY K-MEANS

The k-means method is a widely used clustering technique which minimize the average squared distance between points in the same cluster. Although it offers no accuracy guarantees, it's simplicity and speed are very appealing in practice like that. After many test we thought the best way to initialize is random. We must remember that the number of cluster is going to define the distortion in our model. That is the reason to get and evaluate the distorsion with different number of clusters.

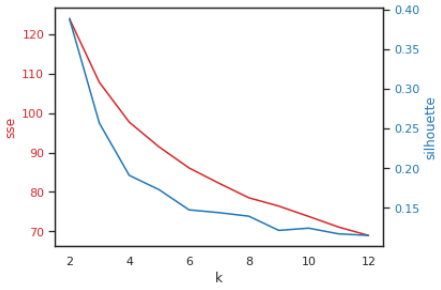

Figure 4: Silouehette and Distorsion

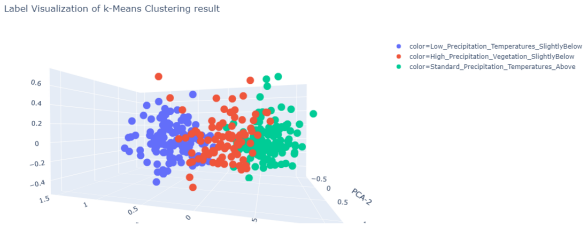

Figure 5: Result from aplicattion of K-Means

Once we got the Silouehette and Distorsion, we select the one that give us the higher Silouehette and the lower Distorsion (Figure 4). We choose 3 like number of clusters, because in our opinion that gives us more information without too many losses. Then the algorithm metrics assigns each register to a group.

After visualization, we should make some representation of the data to assign a label to each group, based on the characteristics of each. Below are shown the labels and descriptions given to each group

- `Group 1 - Low_Precipitation_Temperatures_SlightlyBelow` - Low precipitation and temperatures slightly below average.
- `Group 2 - Standard_Precipitation_Temperatures_Above` - Temperatures above average and standard precipitation.
- `Group 3 - High_Precipitation_Vegetation_SlightlyBelow` - High precipitation, relative humidity and northwest's vegetation slightly below average.

5 HIERARCHICAL CLUSTERING ALGORITHM

The Hierarchical Clustering Algorithm is a typical clustering analysis approach via partitioning dataset sequentially. This algorithm uses distance matrix as clustering criteria, so we had to compute this similarity matrix before executing the algorithm. We can see in figure 4 that this similarity matrix looks like a chess board, due to weeks in data are more similar to other weeks from the same month/season but different year, than to weeks of the same year but different month/season.

Once we got the similarity matrix, we executed the hierarchical clustering algorithm, testing different cluster_distances_measures and plotting the resulting dendrogram. In our opinion, the best

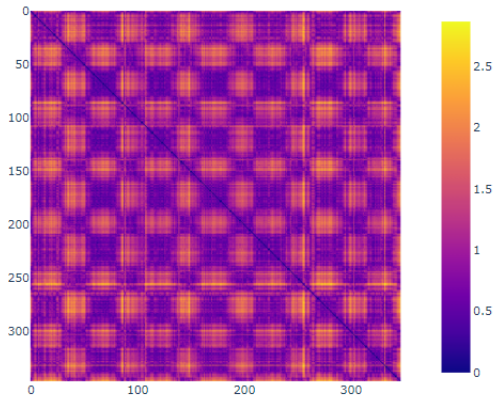

Figure 6: Similarity Matrix of the data

solution was to use 'complete' as linkage criterion as we got concentrated data and this criterion allow us to break up big groups. Using this linkage criterion we got a more balanced dendrogram than using other linkage criterion as single-link or average link.

We've decided to cut the dendrogram resulted of the hierarchical clustering algorithm (Figure 5) by 13, so we get a total of 5 groups, 3 big groups and 2 small groups.

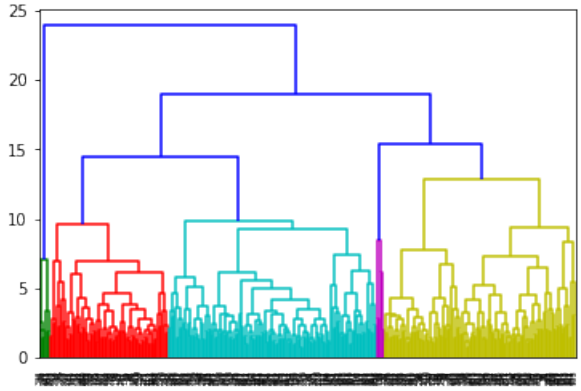

Figure 7: Dendrogram resulting of Hierarchical Clustering Algorithm, cut = 13

After choosing the best dendrogram, we had to characterize the obtained groups in the data. For this task, we made some visualizations which allowed us to assign a label to each group, based on characteristics of each. Below are shown the labels and descriptions given to each group

- `Group 1 - Low_Precipitation_Temperatures_Below` - Vegetation on the west above average, very low precipitation and temperatures below average.
- `Group 2 - Standard_Precipitation_Temperatures_Above` - Temperatures above average and standard precipitation.
- `Group 3 - Standard_Precipitation_Temperatures_ThermalAmplitud` - Vegetation and thermal amplitude below average and standard precipitation/temperature.

- `Group 4 - High_Precipitation_RelativeHumidity` - High precipitation and high relative humidity.
- `Group 5 - Low_Precipitation_RelativeHumidity_Below` - Low precipitation and relative humidity below average.

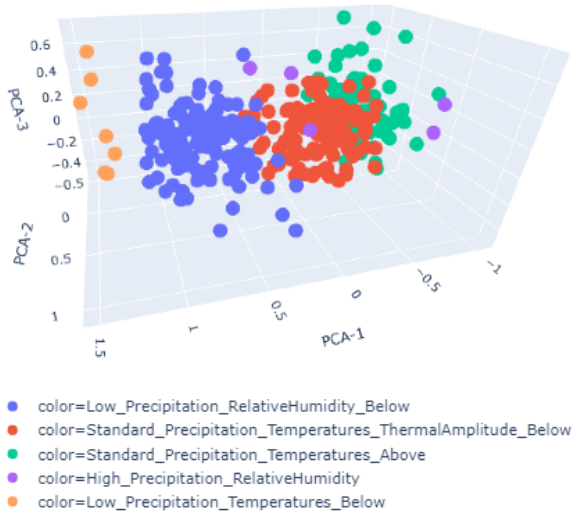

Figure 8: Graphical result of Hierarchical Clustering Algorithm