Machine learning approaches to Reality Mining: The Reality Mining dataset

Elsa Scola Martn

Abstract

In this document I report my proposal for the application of unsupervised classification methods to the Reality Mining problem. Clustering has been the selected method for the discovery of correlation among people who frequently work in the same place or near around. The performed tasks were: preprocess the data set and preparing the data for clustering, then, calculate the optimal number of clusters using the ”elbow” method, to finally compute k means. After that, the data has been visualized after applying PCA (Principal Component Analysis) and the results have demonstrated that cell tower information is not a good indicator of the proximity of subjects that work on the same area.

Contents

1 Description of the problem

The Reality Mining project was conducted at the MIT Media Laboratory. The data set includes information about 106 users location, communication, proximity and activity, which has been collected during 9 months from 2004 - 2005. Among them, 94 subjects participated in the survey including dyadic questions about proximity and friendship with other subjects and some individual habits about social activities.

As there was not a specific task to solve in this project, I had to analyze the data set to extract which was the most relevant question. By observing the data set I came to the conclusion that the most interesting data (from a profitable point of view) is the information about location, as well as proximity and friendship between subjects (results of the survey). The reasoning behind this decision is that it looks like there could be a relationship between this two groups of data and that could also respond to the rising demand in the market of finding patterns in communities, for example to send a certain type of ad to a group of people as it can be inferred that they tent to respond to the same characteristics.

Therefore, the objective of this problem was to use the Reality Mining data set to find correlation among people who frequently work in the same place or near around.

2 Description of the data set

One of the most challenging parts of this project, was understanding the structure of the data set, as it came on a matlab format (.mat) that had many nested arrays. The following breakdown describes only the useful parts of the dataset for this project.

The data set has the following characteristics:

- **'network'** contains the data collected on the online survey completed by 94 of the 106 Reality Mining subjects. The network survey data is divided in 4 parts:
  - *'friends'*: 94 arrays (each one represents a subject of the survey) in which is represented if each person is part of the close circle of friends of the subject. '1' for Yes, 'nan' for No.
  - *'outlab'*: 94 arrays (each one represents a subject of the survey) in which is represented the proximity (within 10 feet/3 metres) with each person outside the lab 5 (at least 4-8 hours per day), 4 (at least 2-4 hours per day), 3 (at least 2 hrs - 30 minutes per day), 2 (at least 10 - 30 minutes per day), 1 (at least 5 minutes) and 0 (default 0-5 minutes).
  - *'lab'*: 94 arrays (each one represents a subject of the survey) in which is represented the proximity (within 10 feet/3 metres) with each person inside the lab 5 (at least 4-8 hours per day), 4 (at least 2-4 hours per day), 3 (at least 2 hrs - 30 minutes per day), 2 (at least 10 - 30 minutes per day), 1 (at least 5 minutes) and 0/default (0-5 minutes)
  - *'subsort'*: an array of 94 numbers (each one represents the number/identifier of a subject on the survey).
- **'s'** includes the subject data of 106 individuals. In this part can be found the data of time-stamped tower transitions among many other groups of data, but we will focus on this set.
  - *'s(n).locs'*: Time-stamped tower transitions. [date, areaID.cellID] (0 is no signal)

We are also provided with the 24 cellular towers that are associated with users work place (according to MITs research, 27 are in the original database, but three pairs of them are duplicated tower IDs) See RealityMining_ReadMe[2] document.

3 Description of my approach

I organized the implementation of the project according to the tasks:

1. Preprocessing of the data set
2. Clustering and visualization
   (a) Apply the ”elbow” method
   (b) Apply K-means
   (c) PCA (Principal Component Analysis) technique application
   (d) Visualize the data
3. Analyze and contrast the results

3.1 Preprocessing

I am interested only in the subjects that completed the survey so in the end I can compare my results to the survey without having missing data. For this reason, I created a function that loops through all the connections to towers registered in each subject and stores in a list the subjects that have been at list once near the towers associated to the workplace (as we are only interested in subjects that have ever been in the workplace). As a result, we obtain the list of subjects, which are 63.

As we will need to compute the time difference between tower connections (to know how much time has been each subject on a specific location in the workplace), we have to convert the dates from matlab format to Python format, so then we can compute the difference.

The goal is to construct a reduced data set that represents the relation of frequency and time spent in each tower for each subject. For achieving this, I created a function that given an empty matrix and a threshold returns in that same matrix the next data:

- each row represents a subject (of the previously selected ones)
- from column 0 to 23 it is represented the time spent in each of the 24 towers that represent the workplace
- from column 24 to 47 it is represented the number of times a subject has been detected from each of the 24 towers that represent the workplace

Note that, we only add time to the time spent in a tower and 1 to the frequency of that tower if the time spent in that tower is greater than the threshold. We do this because if the time spent is less than the threshold we consider that the subject has not been working in that place. I used the pandas library to represent this data as a dataframe.

After computing the frequencies and time spent in each tower, I notice that there are several towers that no subject has connected to, therefore, as they do not add any value, we don’t take them into account. These towers are 7-20, 22 and 23, so we delete their corresponding times and frequencies.

3.2 Clustering and visualization

3.2.1 ”Elbow” method

I will be using the k-means algorithm to solve the problem which is a simple unsupervised machine learning algorithm that groups a data set into a user-specified number (k) of clusters. The algorithm clusters the data into k clusters, even if k is not the right number of clusters to use. Therefore, before using k-means clustering, I needed some way to determine whether they I am using the right number of clusters.

One method to validate the number of clusters is the elbow method. The idea of the elbow method is to run k-means clustering on the data set for a range of values of k (I did it for k from 1 to 20), and for each value of k calculate the sum of squared errors (SSE).

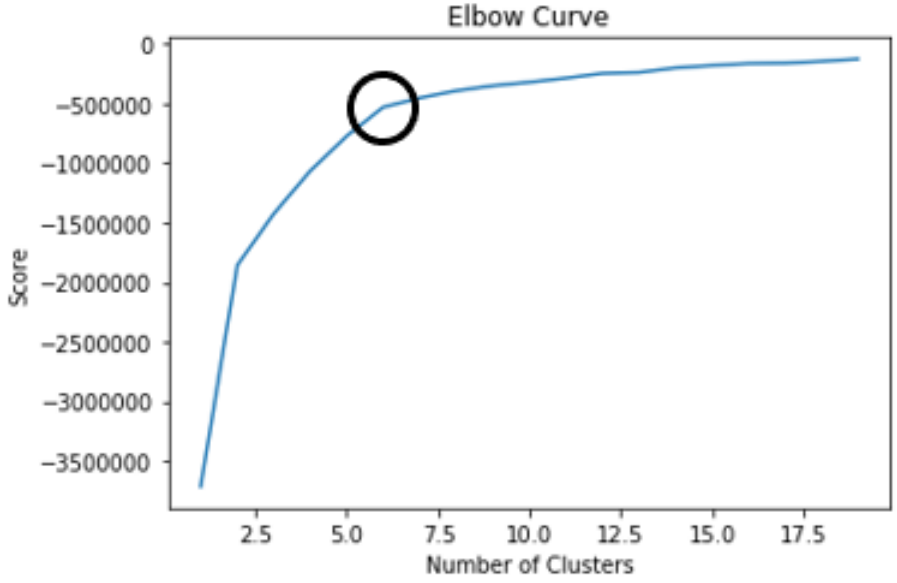

Figure 1: line chart obtained by applying the ”elbow” method

If the resulting line chart looks like an arm, then the ”elbow” on the arm is the best value of k. The idea is that we want a small SSE, but that the SSE tends to decrease toward 0 as we increase k (the SSE is 0 when k is equal to the number of data points in the data set, because then each data point is its own cluster, and there is no error between it and the center of its cluster). So our goal is to choose a small value of k that still has a low SSE, and the elbow usually represents where we start to have diminishing returns by increasing k.

It can observed (see Figure 1) that the optimal value on the ”elbow” is 6. This number is the same that was obtained on a study (by their optimization process) on the ”Reality Mining” data set: K-means based clustering on mobile usage for social network analysis purpose [3], this ensures that this really is the optimal value.

3.2.2 K-means

After analyzing the problem, which consisted in finding correlation among people who frequently work in the same place, I decided that the best method for discovering patterns in a set where there are not any classes was applying clustering, which is used to find out and describe the patterns hidden in the data [4], therefore, clustering is an unsupervised learning task. A cluster is a collection of data objects that are similar to one another within the same cluster and are dissimilar to the objects in other clusters [5].

There are several clustering algorithms that can be used, I decided to apply K-means clustering, which is used when you have unlabeled data, and its goal is to find groups in the data, with the number of groups represented by the variable K. The algorithm works iteratively to assign each data point to one of K groups based on the features that are provided. Data points are clustered based on feature similarity. The results of the K-means clustering algorithm are:

- The centroids of the K clusters

- Labels for the training data (each data point is assigned to a single cluster) subjects that represent the in respect of the frequency and time spent in towers that represent the work-place

3.2.3 Principal Component Analysis (PCA)

The main goal of a PCA analysis is to identify patterns in data; PCA aims to detect the correlation between variables. If a strong correlation between variables exists, the attempt to reduce the dimensionality only makes sense. In a nutshell, this is what PCA is all about: Finding the directions of maximum variance in high-dimensional data and project it onto a smaller dimensional subspace while retaining most of the information.

As the results I have obtained are multidimensional, I cannot visualize them, for this reason it is convenient to use the Principal Component Analysis procedure to reduce it without losing coherence in the data.

After applying this procedure to the data and the previously obtained centroids, I get the new clusters and centroids. This data can now be visualized.

3.2.4 Data visualization

Now the data can be visualized. To add more value to the visualization I compute a function while I visualize each subject in the scatter plot, that for each subject shows the tower in which they spent most of their time at work.

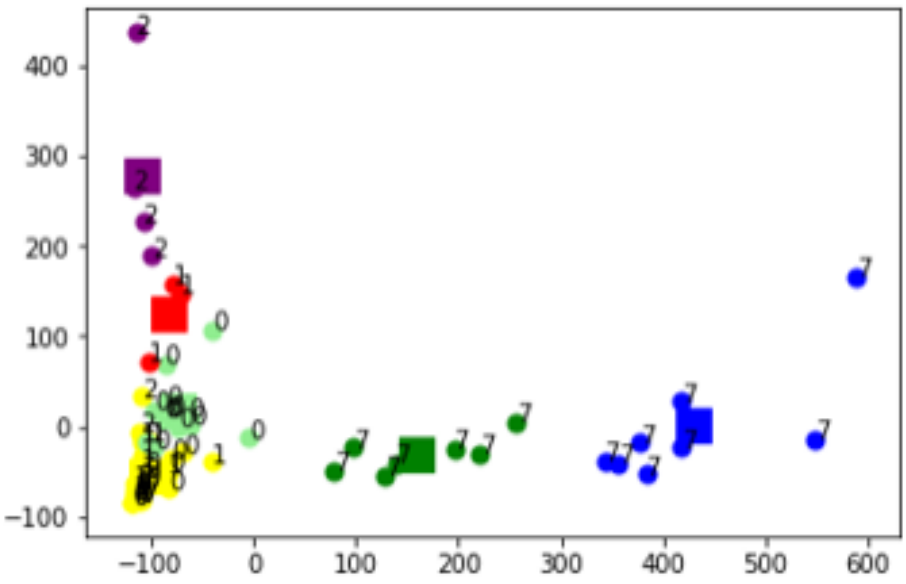

Figure 2: scatter plot with corresponding most frequent tower

It can be observed that there is a correlation between the clusters and the most frequent tower for each subject. Some may ask why the algorithm does not group the clusters with the most frequent tower equals 7 in the same cluster, this is caused by the second most frequent tower of subjects in each of those two clusters, as they differ, they have been separated in two groups.

After this,the same data is visualized but with the subject ID. The results that will be discussed on the next section will have to do with Figure 3.

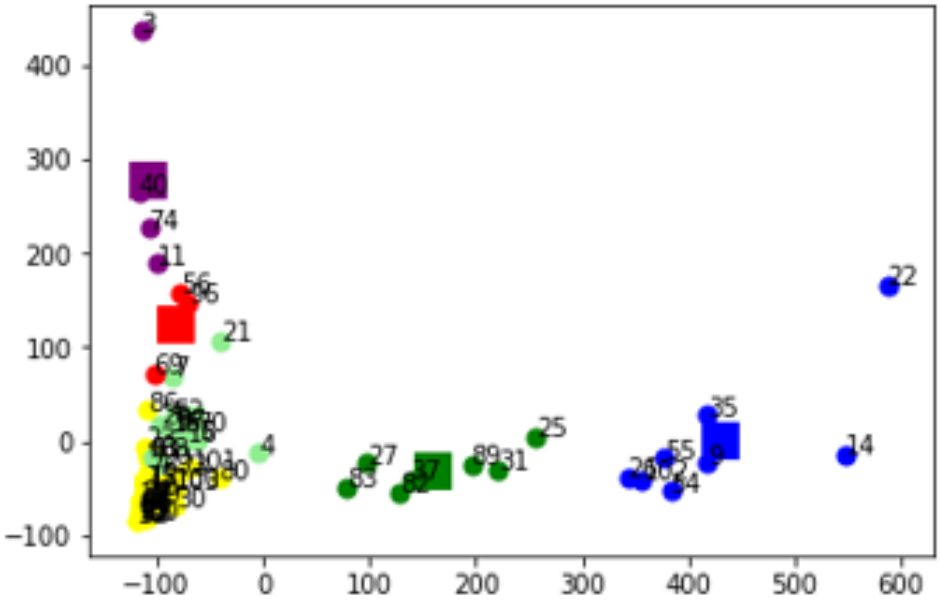

Figure 3: scatter plot with corresponding ID for each subject

3.3 Results

Now my objective is to see if there is any correlation between the location results obtained and the answers of each subject to the 'lab' and 'friends' survey mentioned in the beginning of the document.

For this, we create a function 'groupSubjectInClusters' that groups the subjects in clusters in a matrix when given the clusters and the subjects. This matrix will be used as an input for the 'computeQuality' function, which returns the success rate (between 0 and 1), considering as a success when a subject was valued from another subject of the same cluster with a proximity between 3-5. It is not taken into account results of proximity of 1 or 2 as we previously have used as threshold 1h, for taking into consideration a subject.

The success rate is of 0.026 approximately.

A similar function is created to obtain the friendship success rate, considering as a success when 2 subjects of the same cluster are friends.

The success rate this time is of 0.006 approximately (much worse).

4 Conclusions

Therefore, by seeing the results I came to the conclusion that the frequency and time spent in a tower is not a good indicator of the proximity and friendship (respectively) of subjects that work on the same area. The reason behind these results may be that this subjects, even if they work in the same area, are not so close to interact with each other and foster a friendship as could be working autonomously.

5 Implementation

All the project steps were implemented in Python. I used pandas for reading and preprocessing the dataset, and scikit-learn for the clustering tasks. I illustrate how the implementation works in the Python notebook `Project16-Scola-Notebook.ipynb`.

References

[1] Nathan Eagle, Alex (Sandy) Pentland, Journal Personal and Ubiquitous Computing Volume 10 Issue 4, March 2006 Pages 255 - 268

[2] Nathan Eagle, Alex Pentland, and David Lazer. Inferring Social Network Structure using Mobile Phone Data, Proceedings of the National Academy of Sciences (PNAS), 2009, Vol 106 (36), pp. 15274-15278.

[3] Yang, Xu & Wang, Yapeng Wu, Dan Ma, Athen. (2011). K-means based clustering on mobile usage for social network analysis purpose. 223 - 228.

[4] Hand, D., Mannila, H. and Smyth, P. Principles of Data Mining. Massachusetts Institute of Technology, 2001.

[5] Jiawei, H. and Micheline, K. Data Mining:Concepts and Techniques Second Edition. Morgan Kaufmann Publisher, 2006.