Clustering on emails

Marcos Teixeira
RA 209814
m209814@g.unicamp.br

Miguel Rodrguez
RA 192744
m.rodriguezs1990@gmail.com

1 Introduction

Nowadays, email is one of the most popular forms of communication, mainly due to its efficiency and low cost. In order to explore potential informations in emailing, various data mining techniques have been applied on email data. Clustering emails into smaller groups according to their inherent similarity, facilitates discovering discriminant informations in order to analyze the emails distributions. In this work, we explore some ideas of data mining using a variation of K-means clustering approach on emails. We use a dataset of 19,924 emails to evaluate our performance. A study about PCA approach for this task is also presented, in order to reduce the dimensionality of the feature vectors.

1.1 Dataset

The dataset is composed of 19,924 documents, which are labeled with more than one label for each, with 856 being the total number of labels in the database. In Fig. 1 we can see that the distribution of the labels is very unbalanced, being that the 20 labels with more frequency represent about 60% of the dataset. The average of the frequencies is 37.8 occurrences, standard deviation 154.7, the median 3 and the mode is 1, with this data we can infer that the database is very unbalanced, since more than 50% of the labels only appear once, so finding an optimal number to group this dataset is a very difficult job.

The BOW methodology represent each document as a histogram of occurrences of visual words predefined. In our case, the dictionary is composed by 2,209 words, where each bin represents the probability of the appearance of that dictionary word in the email.

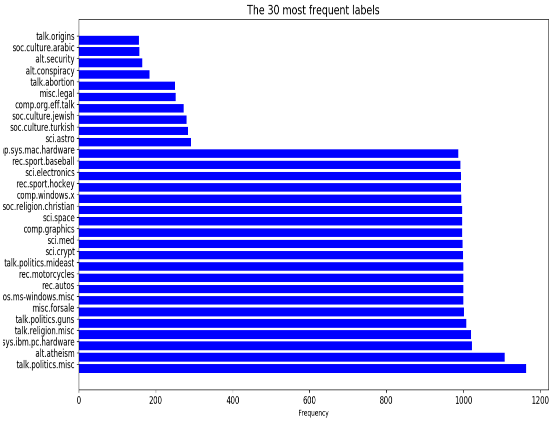


Figure 1: 30 subjects that contain the highest occurrence frequency in the dataset.

2 Material and Methods

2.1 k-Means

k-Means is one of the simplest unsupervised learning algorithms for clustering problem. We first choose K initial centroids, where K is a user-specified parameter, namely, the number of clusters desired. Each point is then assigned to the closest centroid, and each collection of points assigned to a centroid is a cluster. The centroid of each cluster is then updated based on the points assigned to the cluster. We repeat the assignment and update steps until no point changes clusters, or equivalently, until the centroids remain the same.

In this work, we employ a variation of this simple k-means, which is called k-means++ (Arthur, 2007). The difference from k-means is the centroids initialization. K-means++ tries to initialize the cluster as far away from each other as possible. After this initialization, the k-means follows its normal workflow. This method for finding a proper seeding for the choice of initial centroids yields considerable improvement over the standard implementation of the k-means algorithm.

One of the problems of the BOW approach when working with texts is the sparsity of feature vectors. Although our feature vectors have a small value compared to larger problems (e.g. 500,000 dimensions), it also suffers from this problem. Since the K-means++ algorithm uses the Euclidean distance as the standard metric, the distances values are very similar for the docs due to this sparsity. To address this problem, we adopted the cosine similarity, which is a metric that have been the basis for successful document retrieval for over 50 years (Singhal, 2007). The cosine similarity metric is the cosine of the angle between the term vectors and can be expressed as follows:

$$\theta = \arccos \frac{\vec{a} \cdot \vec{b}}{||\vec{a}|| ||\vec{b}||} \tag{1}$$

Cosine Similarity will generate a metric that says how related are two documents by looking at the angle instead of magnitude.

2.2 PCA

A common practice when working with high dimensional feature vectors is 'compact' this information into smaller vector. Principal Component Analysis (PCA) is a effective way to do this task, especially on large datasets. This method works by projecting the original observation data, which may have involved many variables, into a few variables (the principal components).

Given a set of data on n dimensions, PCA aims to find a linear subspace (d < n)-dimensional such that the data points lie mainly on this linear subspace. Mathematically, for a given set of data vectors $x_i, i \in 1...t$, the d principal axes are those orthonormal axes onto which the variance retained under projection is maximal.

3 Experiments and Discussion

In this subsection, we analyze the experimental evaluation results for the methods presented previously.

To choose the ideal cluster number for a given dataset it is necessary to study the behavior of the clusters using different techniques. For our experiments we will use two methods that are complementary to make this choice.

The first method is the election by the study of the Elbow curve, which consists of obtaining each of the centroids and calculating the label for each sample, calculating the average distance of the samples to their respective clusters, and doing this for all the values of k that we want to study. Then the choice of k is based on the graphic generated, and is made through the choice of points where the graph has abrupt changes.

The second way to evaluate the number k is using the Silhouette Coefficient. Although is very subjective the way that we can evaluate our clustering approaches, this metric is very popular for clustering. It is defined as follows:

1. For the i_{th} object, calculate its average distance to all other objects in its cluster. Call this value a_i

2. For the i_{th} object and any cluster not containing the object, calculate the object's average distance to all the objects i the given cluster. Find the minimum such value with respect to all clusters; call this value b_i.

3. For the i_{th} object, the silhouette coefficient is $s_i = (b_i - a_i)/ \max a_i, b_i$

The value of the silhouette coefficient can vary between -1 and 1, where positive represents appropriate separations and negative value generally indicate that a sample has been assigned to the wrong cluster.

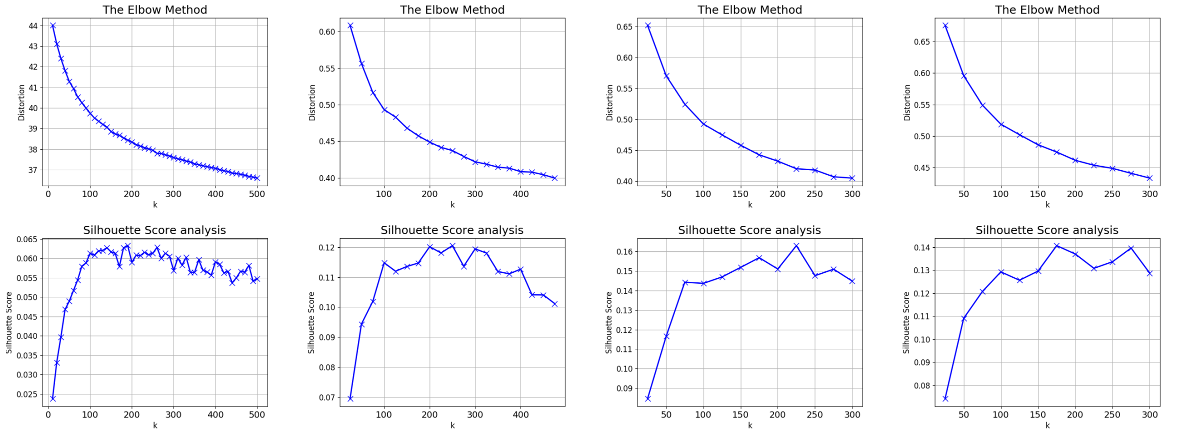
3.1 Experiment 1

In the first experiment, we evaluate a wide range of k, showing the Elbow and silhouette coefficient analysis.

Our first calculations of elbow were made using a k-means with euclidean distance and looking in a range with $10 \leq k \leq 500$, in Fig. 2a we can see the Elbow and silhouette coefficient analysis, where both indicate that the best values of k are in the values 140 and 190, with 140 being the number of clusters that has the best result in the silhouette analysis with a value of 0.0632, which is a very low value.

After analyze the problem, the Euclidean distance does not work very well with sparse data, that is why we decided to conduct a new experiment using the Cosine similarity, because it is well known that this measure of similarity works well with text documents.

In the Fig 2b we can see the Elbow and coefficient of silhouette analysis for k-means using the cosine similarity to perform the clusters. We can



(a) Elbow and silhouette coefficient analysis with euclidean distance and without PCA, in the elbow analysis the abrupt changes were seen for values of k = 60, 140, 250 and in silhouette analysis the highest values were k = 140, 190, 250, with values of coefficient 0.0627, 0.0632, 0.0628 respectively.

(b) Elbow and silhouette coefficient analysis with cosine similarity and without PCA, in the elbow analysis the abrupt changes were seen for values of k = 100, 200, 250 and in silhouette analysis the highest values were k = 100, 200, 250, with values of coefficient 0.1148, 0.1201, 0.1205 respectively.

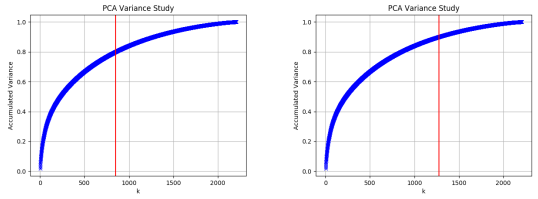
(c) Elbow and silhouette coefficient analysis with cosine similarity and PCA with 80% of variance, in the elbow analysis the abrupt changes were seen for values of k = 100, 150, 225 and in silhouette analysis the highest values were k = 150, 175, 225, with values of coefficient 0.1518, 0.1568, 0.1629 respectively.

(d) Elbow and silhouette coefficient analysis with cosine similarity and PCA with 90% of variance, in the elbow analysis the abrupt changes were seen for values of k = 100, 175, 200 and in silhouette analysis the highest values were k = 175, 200, 275, with values of coefficient 0.1406, 0.1370, 0.1396 respectively.

Figure 2: Elbow and silhouette coefficient analysis using euclidean distance and cosine similarity, and with/without PCA.

see that the values of k that represent possible candidates are 100, 200 and 250, with 250 being the one that obtains the best score in the silhouette coefficient analysis.

3.2 Experiment 2



(a) Number of components with 80% of variance (849).

(b) Number of components with 90% of variance (1274).

Figure 3: Accumulated variance of the PCA study, the red line represents the number of components that represent the required variance.

Considering the large size of the descriptors that represent each document, it is reasonable to think that the performance of clustering algorithms will work better when we reduce the size of our data. In this experiment, we evaluate PCA to reduce the dimensionality of the 2,209-d feature vector. To choose the correct number of components to be used in the reduction by means of PCA it is necessary to carry out a study of the behavior of the variance and obtain an accumulated variance, which tells us what is the number of components necessary to obtain the required variance. In Fig. 3a and 3b, we can see which accumulative variance of the dataset, where the red line represents the number of components that represents the variance of the 80% and 90% which are 849 and 1274 components respectively. After doing the dimensionality reduction, we believe that the candidate values for better k in k-means can vary, this because the data are now less sparse, so we decided to redo the Elbow experiments and silhouette coefficient analysis, for the reduction made with 80% and 90% of energy.

In Fig. 3a we can see the elbow and silhouette coefficient analysis made using a PCA reduction of 80%, where we can see that the best candidate values for k are 150, 175 and 225, which obtained 0.1518, 0.1568 and 0.1629 of silhouette coefficient, values much higher than those obtained in the experiments without PCA.

In Fig. 3b we can see the elbow and silhouette coefficient analysis made using a PCA reduction of 90%, where we can see that the best candidate values for k are 175, 200 , 275, which obtained 0.1406, 0.1379 and 0.1396 of silhouette coefficient, values also higher than those obtained in the experiments without PCA, but lower than those obtained with 80% of variance.

In general, the reduction of components is a very good idea to make when we have very large dimensions and sparse, so making a reduction allows grouping methods to work much better and obtain more realistic results.

3.3 Experiment 3

To see the quality of our clustering we can use quantitative and qualitative measures, in this experiment we propose to measure the quality of our centroids using qualitative measures.

The first step to be able to measure our centroid is to obtain the medoids, which are the closest document to the calculated centroid, in simple terms a medoid is the real centroid of the cluster. Then we proceed to obtain the five closest medoids of each cluster, and we proceed to make a qualitative evaluation in relation to the documents that were obtained.

In general it is expected that the documents that are closest to each other share many common words and ideally belong to the same label, but the BOW technique does not take into consideration the context in which the words are used, so two documents that speak of completely different things but that use very similar words can be considered close with the metrics we use.

Of the results obtained we have two types of results, the first type of result is produced by medoids where their closest neighbors belong to the same label or to labels that talk about very similar topics, eg. the medoid has a tag *sci.space*, and the other five medoids have labels like *sci.med*, *sci.space.shuttle*, *sci.space*, *sci.space* and *sci.space*, so we can see that all have a label that has to do with science, reading the original documents of each medoid, they all talk about something related to science.

Finally, there are the results obtained where the medoid has no relationship with its closest. eg. the medoid has a tag *sci.electronics* and its close ones *talk.politics.mideast*, *soc.veterans*, *alt.military.cadet*, *talk.politics.misc* and *misc.headlines*, the studied medoid talks about electronics and the politicians and the military, here we can see that the clustering was done only by the words used, and not using the context, presenting the great problem of the BoW technique.

To qualify quantitatively in addition to the silhouette coefficient, we decided that for each cluster it is necessary to calculate what is the percentage of documents that belong to the most common label in the cluster, so for each cluster we calculate this percentage and then for all documents we calculate the average 0.5678, median 0.5434, standard deviation 0.2438, the minimum 0.1507 and the maximum 1.0. With these results we can evaluate that our clustering method with k = 225 and PCA with 80 % variance is 56 % effective to clustering the documents of the same label. In general, we can see in the metrics that the clusters have an acceptable clustering, but it can be improved taking into account the contexts of the documents when creating the feature vectors.

4 Conclusions

Problems related with text classification always have to deal with feature representation. In this work, we have initially 2,903 values to describe each document on dataset, being generated by a BoW strategy. Given the sparsity of the vector combined with the number of tags associated for each email and the unbalanced behavior of this tags, this is not a easy task. Furthermore, BoW does not take into account the contexts of documents and words. Therefore the initial simple results of k-means++ has its justification.

After the observation that euclidean distance for k-means++ is not the best choice for text, we change to cosine similarity that is commonly used for this problems. The results showed an significantly improved but the dimensions of the feature vectors also complicates the algorithm. Using PCA we obtained the higher silhouette coefficient (0.163) with k=225, using 80% variance of the data. However the results of our qualitative study shows that there some cases where the neighbors of the medoids are not talking about the same subject.

References

Arthur, David, and Sergei Vassilvitskii. 1997. *k-means++: The advantages of careful seeding.*. Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms. Society for Industrial and Applied Mathematics.

Singhal, Amit 2001. *Modern Information Retrieval: A Brief Overview.* Bulletin of the IEEE Computer Society Technical Committee on Data Engineering 24 (4): 3543.