CS 565 Data Mining, Fall 2017 | Instructor: Prof. Terzi Evimaria

Project 1

Nov 04th, 2017 | Tian Zhang

1. Description

In this project, we will explore clustering and experiment with high-dimensional data. We will be using a movies dataset. In this dataset, we will find data corresponding to 4,803 movies and we try to implement popular algorithms like k-means and k-means++ clustering algorithms efficiently in order to discover the hidden structure. Meanwhile, we also try to use Principal Component Analysis(PCA) to do dimentional reduction.

2. Implementation

(1) Data Prepossessing: loading 'movies.csv' to a dataframe, creating other two extra dataframe to store numerical features and categorical features. For numerical, we used $Result = \frac{realData - minValue}{maxValue - minValue}$ to normalize them. Meanwhile, for categorical feature, because they were storing in JSON string, I tried to parsing them out for JSON string and transfered them to dummy variables(using boolean vector to present categorical features). Finally, I merged numerical dataframe with categorical dataframe.

(2) Clustering: Used Kmeans and Kmeans++ to do clustering for both numerical dataframe and final resulted dataframe got from merging. And comparing their result for modifying data prepossessing strategy(like using Jaccard Distance to measure the distance between original categorical data points.) Meanwhile, tried to draw error curve to find the best value for k

(3) Used scikit-learn to do Principal Component Analysis and tried to visualize the clusters in 2-D plane.

3. Code Usage

python source.py -d <path to file> -c <clusterNum> -init <method to use> ("random" for Kmeans; "k-means++" for Kmeans plus plus.)

4. Result

(1) The value of 'k':

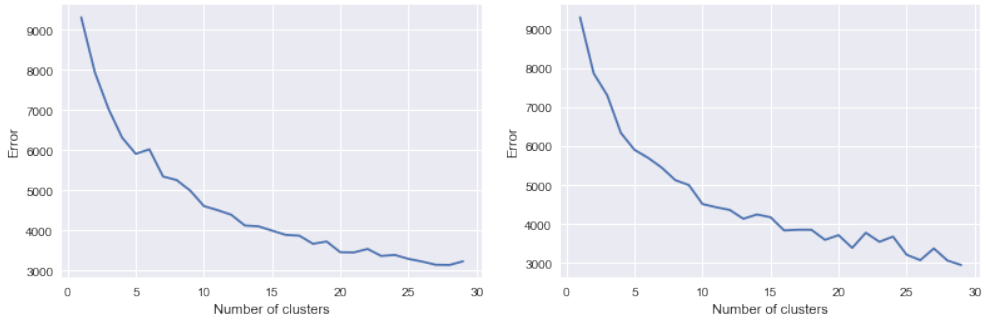

Figure 1: Cost Function Curve for Kmeans and Kmeans++

To find out the suitable 'k' for clustering, I tried to calculate the $\sum$ of distance between each data points and their corresponding centriod over a range of $k \in [1, 30]$, and I got the above two curve for both kmeans and kmeans++. We could find out that when $k = 10$ the cost function decrease rapidly, which means separating the movie data pts to 10 clusters is a reasonable choice.

(2) Principal Component Analysis:

In this part, I used the PCA function from scikit-learn package to reduce high dimensional movie dataset to 2 components vector. Using kmeans to cluster our top 250 movie dataset with a reasonable k(k=5), I got the following two clustering visualization:(From to Figure 2, it is reasonable to set $k = 5$ (Five clusters).

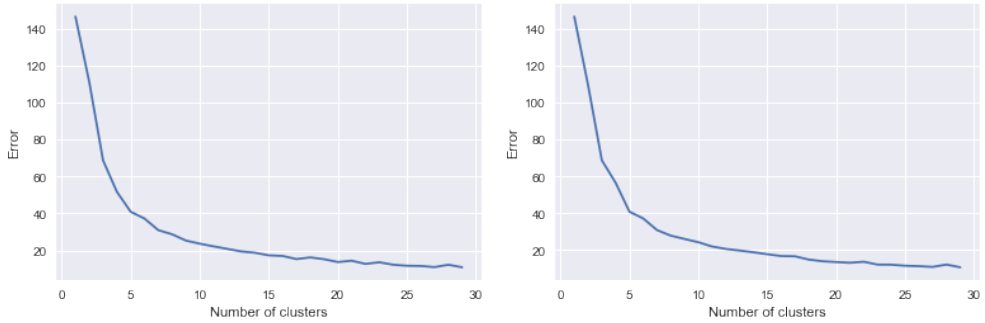

Figure 2: Cost Function Curve for Kmeans and Kmeans++ for top 250 movies

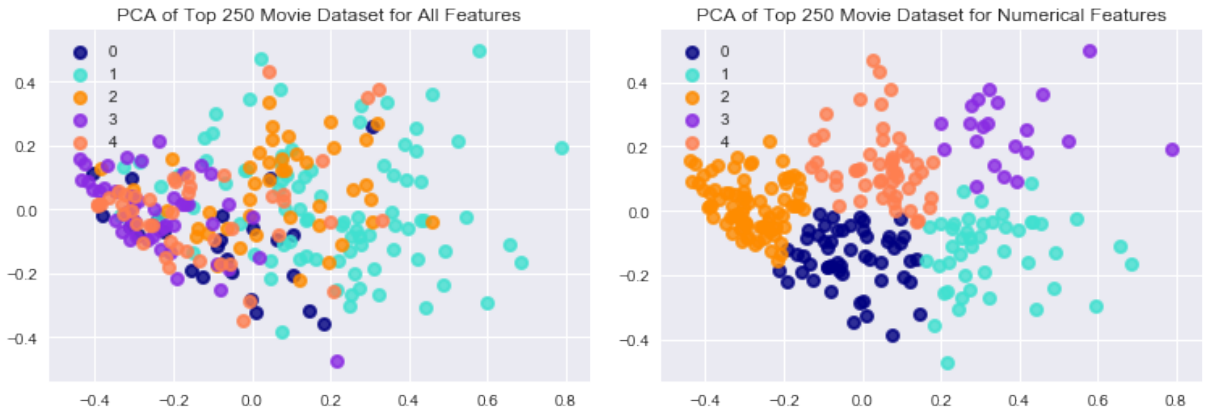

Figure 3: Cost Function Curve for Kmeans and Kmeans++

Comparing the above two clustering results, it is obviously that only using numerical features performed better than using all prepossessed dataset, which leads me to modify my method for transferring categorical data to real value vectors.

5. Future Work

- Using other method to transfer text to vectors(like word2vec, embedding text to fixed dimensional vectors, including n-gram, skip-gram and so on)
- Trying to modify the new centriod selection rather than only choosing mean data point.
- After reading some paper, I found out that we can use other strategies to find the reasonable k suit our target dataset. (like Silhouette Coefficient, Calinski-Harabasz Principal and Duda-Hart test.)