George Mason University
Department of Computer Science 584-006
Virginia,USA

K Means Clustering Analysis

Abhishek Jallawaram, Ragnarok7417, 0.95 (iris), 0.76 (image)
ajallawa@gmu.edu, G01373042

Abstract

This document is provides process of handling image datasets using feature selection, dimensionality reduction and selecting the best clustering model using K-Means algorithm.

1. Introduction

This document is covers implementation of non-linear dimensionality reduction techniques like T-distributed stochastic neighbor embedding (t-SNE), Uniform Manifold Approximation and Projection for Dimension Reduction(UMAP) and K-Means clustering algorithm on an image dataset.

1.1. Objectives

Listed below are the objectives achieved in this assignment:

- Dimensionality Reduction techniques
- Implement the K-Means Algorithm
- Evaluation Metrics for clustering - Sum of Squared Error (SSE)
- Choose the best model i.e., best initial random centers

2. Problem Scenario

2.1. Iris dataset

Assign the dataset into 3 clusters.

2.2. Image dataset

Assign the dataset into 10 clusters
Build best K-Means clustering model which has the best feature engineering performed, best initial random centers picked and evaluation performed to attain the best results.

3. Feature Engineering

The below linear and non-linear dimensionality reduction methods were deployed on the image dataset to reduce the number of features to an optimal amount for better visulaization and computation.

- Principal Component Analysis(PCA)
- Singular Value Decomposition(SVD)
- t-distributed Stochastic Neighbor Embedding (t-SNE)
- Uniform Manifold Approximation and Projection for Dimension Reduction (UMAP)

3.1. Linear Dimensionality Reduction

Linear dimensionality reduction techniques like Principal Component Analysis (PCA) and Singular Value Decomposition(SVD) were performed on the image dataset to efficiently reduce the features using the variance. We have observed that linear dimensionality reduction techniques were these algorithms define specific rubrics to choose an “interesting” linear projection of the data. These methods can be powerful, but often miss important non-linear structure in the data. We have implemented the below non-linear dimensionality reduction techniques like t-SNE, UMAP and MDS to retain this property while performing feature engineering.

3.2. Non - Linear Dimensionality Reduction

3.2.1. t-SNE

t-distributed Stochastic Neighbor Embedding (t-SNE) converts affinities of data points to probabilities.The affinities in the original space are represented by Gaussian joint probabilities and the affinities in the embedded space are represented by Student’s t-distributions.

t-SNE will focus on the local structure of the data and will tend to extract clustered local groups of samples. This ability to group samples based on the local structure might be beneficial to visually disentangle our dataset which comprises several manifolds at once.In our scenario, the default metrics were used to transform the data into 2 dimensions.

3.2.2. UMAP

UMAP, at its core, works very similarly to t-SNE - both use graph layout algorithms to arrange data in low-dimensional space. UMAP constructs a high dimensional graph representation of the data then optimizes a low-dimensional graph to be as structurally similar as possible.

In order to construct the initial high-dimensional graph,UMAP builds something called a ’fuzzy simplicial complex’. This is really just a representation of a weighted graph, with edge weights representing the likelihood that two points are connected. To determine connectivity,UMAP extends a radius outwards from each point, connecting points when those radii overlap. Choosing this radius is critical - too small a choice will lead to small, isolated clusters, while too large a choice will connect everything together.UMAP overcomes this challenge by choosing a radius locally, based on the distance to each point’s nth nearest neighbor.UMAP then makes the graph ’fuzzy’ by decreasing the likelihood of connection as the radius grows. Finally, by stipulating that each point must be connected to at least its closest neighbor,UMAP ensures that local structure is preserved in balance with global structure.

In our scenario, we hvae used the default metrics to reduce the dimnesionality of data into 2 dimensions.

4. K-Means Algorithm

The initial cluster centres are determined by selecting **K** number indices at random using random seed and random sample. A proximity measure like L1-Norm (Manhattan Distance), L2-Norm (Euclidean Distance) or similarity measure like cosine similarity is used to calculate the distance/simalarity between the initial clusters and all the datapoints in the dataset.

The datapoints are allocated to different clusters based on the minimum distance or maximum similarity. The cluster centers of next iteration are determined by taking the mean of the samples in the cluster. The iterations are performed till the maximum limit is reached(10000) or we attained the cluster centers which cannot be further processed i.e clusters centers of the next iteration calculated equal the current iteration cluster centers.

The model is evaluated using Sum of Squared Errors by calculation the errors in each cluster. We take the SSE for each datapoint with it's respective cluster center. The total SSE is calculated by adding the individual SSE of each cluster.

The algorithm is run 20 times with different initial random cluster centers and the SSE is calculated. The model with least SSE is used to determine the final clusters.

Another variant of this algorithm is by using the median to calculate the cluster centers for the next iteration (K-Medoids).In our scenario, we have implemented euclidean distance, manhattan distance and cosine similarity with K-Means and K-Medoids algorithms using SSE as the evaluation metric to attain the best model.

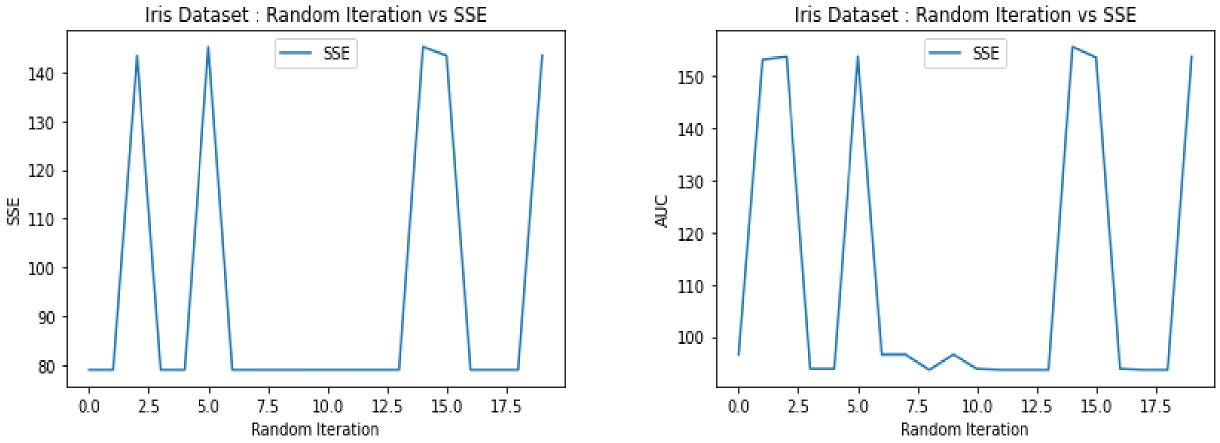
5. Obervation

5.1. Iris dataset

We have attained the best results when cosine similarity as similarity metric was used on the iris dataset with both K-Means and K-Medoids as the algorithm. The algorithm was run 20 times with different initial random clusters and SSE was used as evaluation metric.The model with minimum SSE was used to achieve the best results.(0.95).

5.2. Image dataset

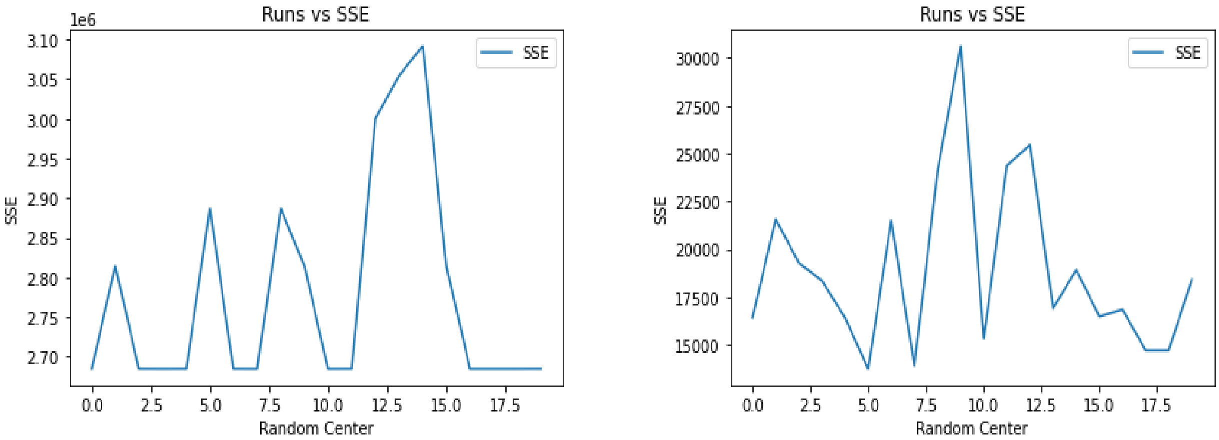
We have implemented the non-linear dimensionality reduction techniques like t-SNE and UMAP and implemented the K-Means algorithm with euclidean distance as the proximity measure. The algorithm was run 20 times with different initial random clusters and SSE was used as evaluation metric.The model with minimum SSE was used to achieve the best results.(0.76)



1 – SSE for Iris dataset with Euclidean distance 2 – SSE for Iris dataset with cosine similarity

Fig 1 – SSE for Iris dataset

Elbow method is one of the most popular method used to select the optimal number of clusters by fitting the model with a range of values for K in K-means algorithm. Elbow method requires drawing a line plot between SSE (Sum of Squared errors) vs number of



1 – SSE for image dataset with t-SNE **2 – SSE for image dataset with umap**

Fig 2 – SSE for Image dataset

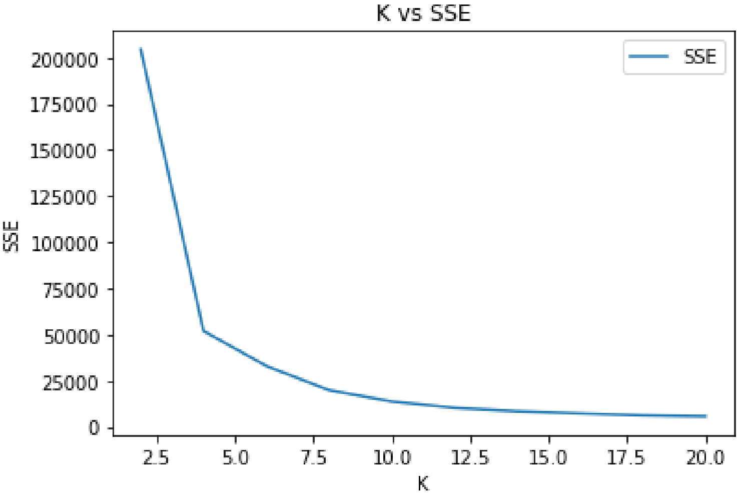


Fig 3 – Image dataset : K vs SSE

clusters and finding the point representing the “elbow point” (the point after which the SSE or inertia starts decreasing in a linear fashion).

6. References

- TruncatedSVD=`https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.TruncatedSVD.html`
- PCA=`https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html`
- SMOTE=`https://imbalanced-learn.org/stable/references/generated/imblearn.over_sampling.SMOTE.html`