COMP5318 - Machine Learning Assignment 2 Report

Zhecheng Zhong(SID:490319299), Xiaoqi Huang(xhua7314)

Semester1 2020

1 Data Setup

1.1 Travel Reviews Data-set

The travel review data set reviews on destinations in 10 categories. It contains 11 attributes. The data pre-processing manipulates columns, separating between the first column and the other ten columns. The first column is stored as userID series while the remaining columns consists trip_advisor matrix.

1.2 ICMLA

In this section ICMLA data is separated and transferred into other types of subset data. According to Deigo[1] ICMLA is split into three contents part — which are keywords, abstract,and title. Via importing NLTK, Tfidfvectorizer packages, the contents is finally changed from string to vector matrix. Paper Matrix = 1DM(abs)+2DM(keywords)+3DM(title), where all the DMs are dissimilarity matrix. Each element in DM(keywords) and DM(title) is the dissimilarity between two samples calculated via Jaccard coefficient. DM(abs) is retrieved through cosine similarity. The gold cluster part would then be determined as standard label as integer value.

2 K-means Clustering

2.1 Travel Reviews Dataset

2.1.1 Similarity measure selection

K-Means is an unsupervised machine learning algorithm that groups data into k number of clusters. Since kMeans clustering algorithm uses only Euclidean distance metric, we re-define another k-means function using cosine distance. After executing program, the elbow method depicts the inertia index. Inertia calculates the sum of squared distances from each point to its assigned center. Figure1 compares the inertia difference between euclidean and cosine similarity. When k¡6, cosine similarity measurement

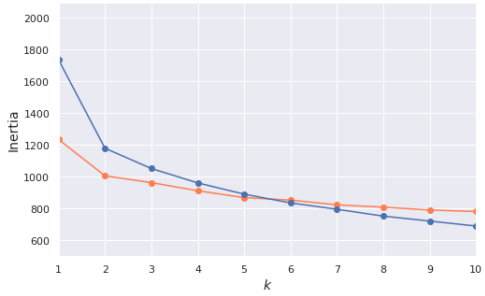


Figure 1: inertia with different clusters. (Blue: euclidean, Coral: cosine)

is more effective; when k¿=6, euclidean similarity measurement is better.

2.1.2 The optimal k selection

we cannot simply take the value of k that minimizes the inertia, since it keeps getting lower as we increase k. By analyzing inertia curve, one can pick up elbow point as an optimal k results. Let's define if Ki is the elbow, then inertia difference between Ki, Ki+1 should be bigger than Ki, Ki-1. Based on figure 1, here we got elbow should between k=6,7,8.

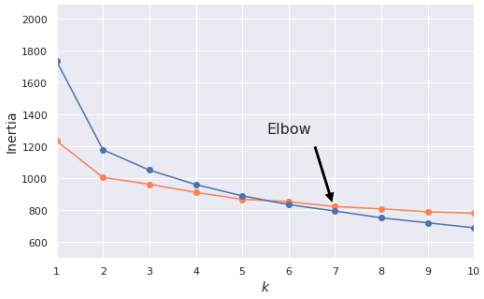


Figure 2: Elbow point, k = 7)

2.1.3 Cluster Evaluation

Another evaluation method used here is Silhouette score. The Silhouette Coefficient is calculated using the mean intra-cluster distance and the mean nearest-cluster distance for each sample. Silhouette coefficients (as these values are referred to as) near +1 indicate that the sample is far away from the neighboring clusters. Figure 3 reveals a proper k should be 6 or 8, and Figure 4 provides more details on the Silhouette Coefficient, giving an insight that when k=6, the whole cluster tends to be more consistent.

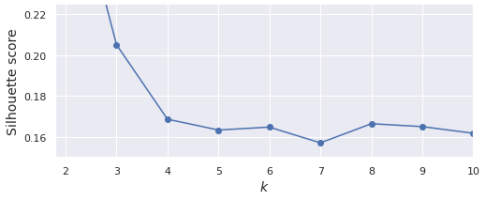


Figure 3: Silhouette score)

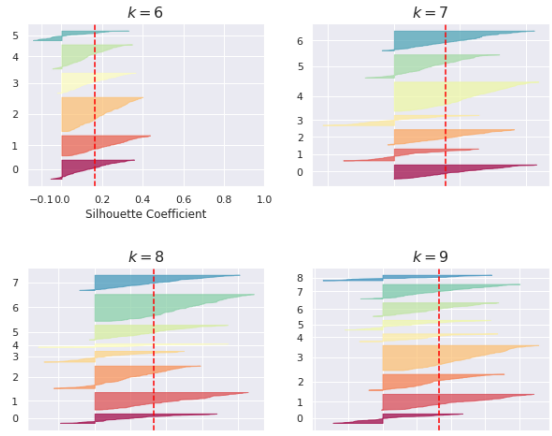


Figure 4: Silhouette Coefficient)

2.2 ICMLA

2.2.1 Similarity measure selection

We utilize similar method computing a set of k-mean with different number of cluster both by using euclidean and cosine similarity methods. Since ICMLA has gold attribute, the gold attribute should be interpret first, then all the predict cluster are measured and compared. Figure 5 draws the purity results with different Ks. Coral line stands for euclidean while blue line stands for cosine. The euclidean similarity tends to be more consistent and with a higher purity.

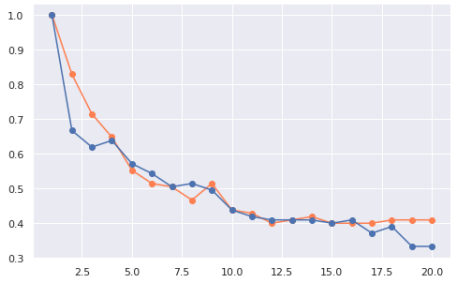


Figure 5: ICMLA: purity with different k(Coral: euclidean, Blue: cosine)

2.2.2 The optimal k selection

Literally, the original gold attributes contain 20 cluster, indicating the proper classification should produce approximate 20 aim cluster. However, the purity calculation algorithm implemented in the project has drawbacks. Two clusters could result in having the same label as long as they have the same most frequent element. Still, according to figure 5, while

we using euclidean, only considering K¿10, the optimization of K is either 11 or 14.

2.2.3 Cluster Evaluation

Here, since the data set offers gold cluster, Normalized Mutual Information based on entropy is wield for evaluation. Normalized Mutual Information tells us the reduction in the entropy of class labels that we get if we know the cluster labels. Figure 6 points out the optimal number of cluster should be 20 which has NMI around 0.45.

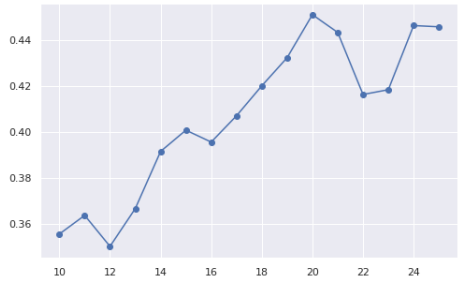


Figure 6: ICMLA: NMI of different cluster number

3 Hierarchical Clustering

3.1 Travel Reviews Dataset

3.1.1 Similarity measure selection

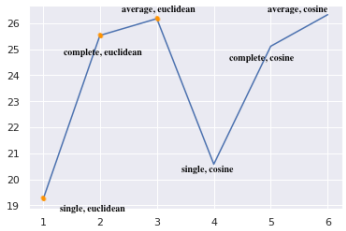


Figure 7: trip advisor: calinski harabaz score

During the hierarchical cluster training, six models are built with different similarity metrics and methods. They are evaluated based on Calinsk Harabaz Score. As a consequence, cosine similarity with different method is generally better than euclidean.

3.1.2 Best distance metrics

Figure 7 suggests that the best distance metrics for trip advisor review is "average"

3.1.3 Cluster Evaluation

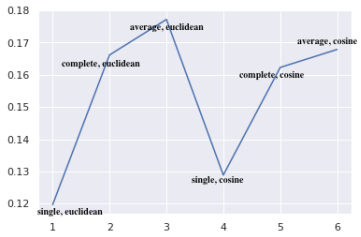


Figure 8: trip advisor: Silhouette

Another cluster evaluation method implemented is Silhouette score. Similarly, this has the semi-shape of Calinsk Harabaz Score in last figure. However, here, it suggests euclidean similarity is more fittable than cosine and euclidean with average methods has the best model. And the implementation results is drawn by figure 9.

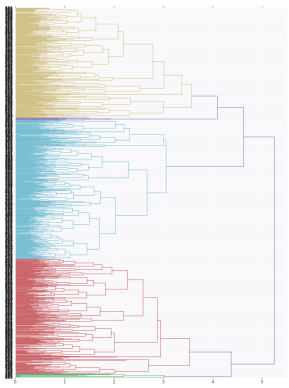


Figure 9: trip advisor: Hierarchical clustering

3.2 ICMLA

3.2.1 Similarity measure selection

Same implementation methods are used in ICMLA, here the hierarchical cluster results and methods comparison curve are in figure 10 and figure 11. They suggests that euclidean complete would be the best offer.

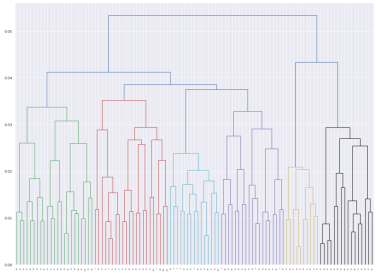


Figure 10: ICMLA: Hierarchical clustering

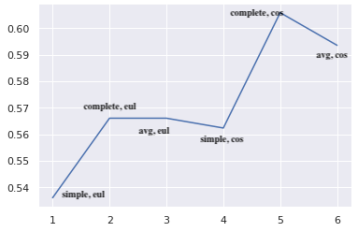


Figure 11: ICMLA: Hierarchical clustering NMI evaluation

3.2.2 Best distance metrics

According to figure 11, best distance metrics is complete.

3.2.3 Cluster Evaluation

Another evaluation executed here is purity, however, due to the shortcomings of the purity calculation function, this evaluation result is not practical.

4 DBSCAN Clustering

4.1 Travel Reviews Dataset

4.1.1 The optimal min_points, and epsilon

In order to determine the optimal minimum points and epsilon index, we create a set of DBSCAN model with different ranging of eps, and min_sample. We must provide a value for epsilon which defines the maximum distance between two points in a cluster. In layman's terms[2], we find a suitable value for epsilon by calculating the distance to the nearest n points for each point, sorting and plotting the results. Then we look to see where the change is most pronounced and select that as epsilon. Hence, we set eps

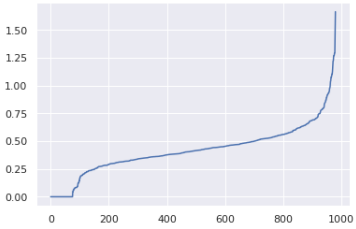
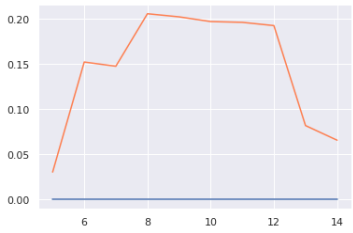


Figure 12: Optimal value for epsilon is found at the point of maximum curvature



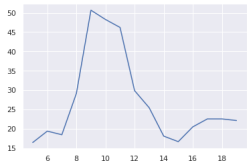
(0.6,0.75), and drawing the curve of silhouette score with different min samples, finding that min_sample = 8

4.1.2 Similarity measure selection

Similarity measurements are euclidean, manhattan and cosine. When eps = 0.7, min_sample ranges in

(5,15), the manhattan and cosine methods tends to have a bad performance. So we will use euclidean.

4.1.3 Cluster Evaluation



As a result, based on the calinski harabasz score evaluation and silhouette score evaluation, the best eps is around 0.7 with minimum sample equals to 9.

4.2 ICMLA

4.2.1 The optimal min_points, and epsilon

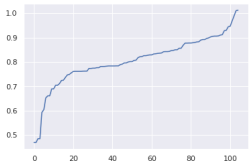


Figure 15: eps = 0.9 brings about the elbow

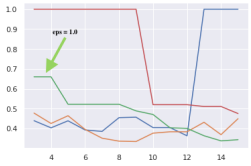
Figure 16: blue line: euclidean, coral line: cosine

Using the same method, the elbow appears when eps = 0.9 and according to the both adjusted rand score and homogeneity score figure, the optimal minimum sample should be 5.

4.2.2 Similarity measure selection

Similarity measurements are euclidean, manhattan and cosine. When eps = 0.9, min_sample ranges in (5,20), the manhattan and cosine methods tends to have a bad performance. The adjusted rand scores are zero. So we choose euclidean.

4.2.3 Cluster Evaluation



According to the figure, it shows when eps 1.0, min score equals to 5 has the best homogeneity score performance.

4.3 The Best Model

4.3.1 Travel Reviews Dataset

Travel Reviews Dataset has its best DBSCAN model, when eps = 0.9 and min_samples = 9.

4.3.2 ICMLA 2014 Accepted Papers Dataset

ICMLA 2014 Accepted Papers Dataset has its best hierarchical model, when metric = cosine and method = complete.

References

[1] Cèsar Ferri Diego Valle jo Huanga, Paulina Morillo. *Semi-Supervised Clustering Algorithms for Grouping Scientific Articles.* International Conference on Computational Science, 2017.

[2] Imas Sukaesih Sitanggang Nadia Rahmah. *Determination of Optimal Epsilon (Eps) Value on DBSCAN Algorithm to Clustering Data on Peatland Hotspots in Sumatra.* IOP Conference Series, 2016.