University of Glasgow

Machine Learning & Artificial Intelligence for Data Scientist M (COMPSCI5100)

Case Study:
Model Selection for Clustering

Student: Ching Hsuan Lin
Student ID: 2702329L

Student: Weijian Ning
Student ID: 2821551N

Table of Contents

1. Introduction

Clustering, as an unsupervised learning task, has the primary objective of learning from unlabeled training samples and dividing different samples in a dataset into several independent subsets based on specific correlations of features between samples. The subsets with the same features are generally referred to as a “cluster".

Model selection for clustering is challenging. First, there is no common standard for the definition of correlation between samples in different clustering models. The choice of different correlation criteria (e.g. distance in KMeans, hierarchical structure in Hierarchical Clustering, the density of HDBScan, etc.) and the distribution form of sample points in different feature spaces (e.g. feature dimensionality, the regularity of cluster shape, etc.) all have an impact on the clustering results.

Second, unlike classification problems, there are usually no reliable truth labels in practical clustering problems to use as a criterion for evaluating how well the clustering works. It also brings the challenge of evaluating the clustering model's performance.

Finally, the understanding of the clustering outcomes is subjective and abstract. Each cluster class may have some implicit common concepts or characteristics unknown to the algorithm in advance. A qualitative analysis of the samples' actual situation is often necessary to understand the semantic concepts corresponding to these cluster classes.

The datasets used in this paper were derived from two different deep neural network-based representations, PathologyGAN(PGE) and Resnet50, extracted from colorectal tissue patches. Each representation size was reduced to 100 using PCA and UMAP dimensional reduction methods.

This paper will first introduce the clustering algorithms and evaluation metrics used in the clustering process and then illustrate the optimisation methods for selecting hyperparameters during the implementation of the model with the specific experimental data results. Next, comparing and evaluating the performance of the clustering models were done through quantitative metrics and qualitative analysis. Finally, a suggestion for the selection of clustering models is given with the evaluation results.

2. Methods Introduction

2.1 K-Means Clustering

K-means Clustering is one of the most accessible unsupervised learning, meaning there is no target or class contained within the datasets. To be more concise, unsupervised learning can predict or cluster an unlabeled dataset. K-means is an iterative solution clustering analysis algorithm. Firstly, the number of clusters must be chosen by how many clusters we need from the datasets eventually, and K points from each set will be randomly selected as K initial clustering centroids. Secondly, calculate the distance between each centroid and each other point in the dataset and assign every point to the closest cluster. Thirdly, recompute the centroids of newly formed clusters, repeatedly estimate the distance between each centroid and each other point and assign them to the nearest clusters until the system converges to a point where assignments do not change. The termination conditions can be that no (or minimum) number of points are reassigned to different clusters, no (or minimum) cluster centroids change again, and the sum of error squares is locally minimum.

2.2 Hierarchical Clustering

Hierarchical Clustering is also an unsupervised clustering method. The biggest difference from K-means is that it can illustrate the relevance between each point or cluster. There are several ways to approach this objective; the one used in this report is Agglomerative Hierarchical Clustering, which is an iterative algorithm. First, every point in the dataset is assigned to an individual cluster, and the distance between each cluster will be evaluated. Second, the separated clusters will be merged with their closest neighbour, and the distance between new clusters can be recalculated. The iteration continues until there is one global cluster left or K clusters (chosen) left.

2.3 Hierarchical Density-Based Spatial Clustering

DBSCAN(Density-Based Spatial Clustering of Applications with Noise) is a density-based clustering algorithm[1]. Unlike partition and hierarchical clustering methods, it defines the cluster as the maximum set of density-connected points, which can divide areas with sufficiently high density into clusters and find clusters of arbitrary shapes in noisy spatial databases. Two main parameters are used in DBSCAN: Eps and MinPts. Eps is a selected distance that can be used to measure a point with one another in the same cluster or not; MinPts is the minimum number of points within a region that can be considered as a cluster. First, one point is selected as the starter, and if there are MinPts points in the distance of Eps from the point, these MinPts points will be assigned to one cluster. The iteration continues until every point in the dataset is visited, and the points not assigned to any cluster will be considered noises.

HDBSCAN(Hierarchical Density-Based Spatial Clustering of Applications with Noise) is a clustering algorithm developed by Campello, Moulavi and Sander[2]. It extends DBSCAN by converting DBSCAN into a hierarchical clustering algorithm and then uses the extraction plane clustering technique based on clustering stability. The biggest difference between HDBSCAN and traditional DBSCAN is that HDBSCAN can handle clustering problems with different densities. There are also two important parameters in HDBSCAN, which are min_cluster_size and min samples. The former determines the smallest size grouping that you wish to consider a cluster; the latter provides a measure of how conservative you want your clustering to be. The bigger the min_cluster_size, the fewer clusters will be generated; the bigger the min_samples are, the more conservative the clustering – more points will be considered as noise, and clusters will be restricted to progressively more dense areas.

2.4 Silhouette Coefficient

The Silhouette Coefficient is an evaluation metric often used in clustering and is usually applied where the truth label information is unknown. Silhouette scores can indicate the goodness of a clustering model by evaluating the mean distance between points within a cluster and the distance between clusters. The silhouette coefficient has a value range of [-1 1]. The closer the distance between samples of the same category and the farther the distance between samples of different categories is, the higher the score becomes.

2.5 V-measure

V-measure is applied when the truth label is given. V-measure scores illustrate whether the clustering output is good by calculating the harmonic mean of homogeneity and completeness. Homogeneity indicates the purity of the output of a model, a well-developed model should predict objects within one cluster belonging to the same label; completeness shows how well the data points with the same label are grouped by the same cluster. The score is a measure between [0 1], which has a better effect when close to 1.

2.6 Adjust Rand index

The Adjust Rand Index is an algorithm for calculating the similarity of two sequences. To be more concise, it evaluates the similarity of two sets of labels by measuring the proportion of agreement between truth labels and clusters. Its value range is [-1 1], and the higher the value is, the more similar the two label sets are.

3. Model Implementation and Clustering Result

3.1 Hyper-parameter optimisation

• 3.1.1 Grid Search

The function of Grid Search can be automated hyper-parameter tuning. The optimised results and parameters can be evaluated as long as the parameters are input. In grid search, parameters are searched, that is, parameters are adjusted in turn according to the step size within the specified parameter range, and the modified parameters are used to train the learner to find the parameter with the highest evaluation metrics score from all the parameters, which is a process of training and comparison. In this report, we are using V-measure to evaluate model performance since the ground truth label is given in the dataset, which is a completive way to evaluate models by calculating the harmonic mean of homogeneity and completeness.

• 3.1.2 K-Means (Grid Search)

In K-means, we grid search from 9 clusters to 30 clusters within integer steps and find the number of clusters with a maximum v-measure score in different datasets, as shown in Table 1. And in Figure 1, we can see how the cluster number affects the V-measure score in different datasets, where the v-measure score is on the y-axis, and the number of clusters is on the x-axis; the red dot is the highest score.

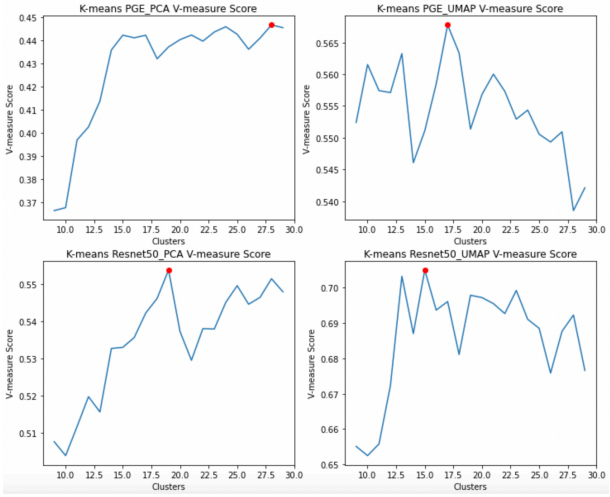


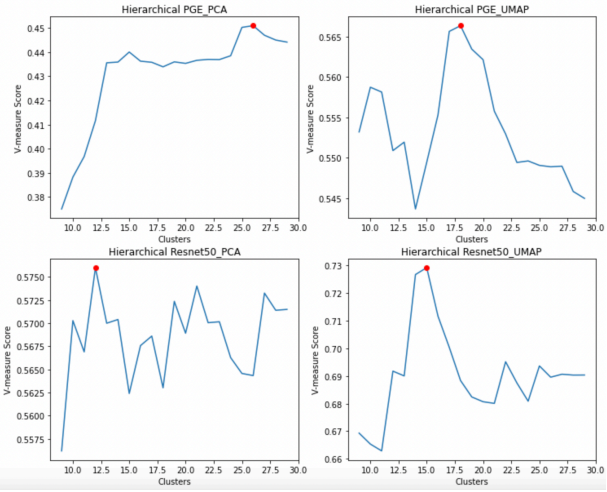
Figure 1. V-measure score of K-means model.

Kmeans	PGE PCA	PGE UMAP	Resnet50 PCA	Resnet50 UMAP
V-measure Score	0.446687	0.567786	0.553885	0.705007
Cluster Numbers	28	17	19	15

Table 1.Grid search result of K-means model

• 3.1.3 Hierarchical Clustering (Grid Search)

In hierarchical clustering, grid search is also set in the range of 9 to 30 clusters within integer steps. The results are shown in Figure 2 and Table 2.



	PGE PCA	PGE UMAP	Resnet50 PCA	Resnet50 UMAP
Heirarchical Clustering				
V-measure Score	0.450976	0.56636	0.576	0.729171
Cluster Numbers	26	18	12	15

Figure 2. V-measure score of Hierarchical model. Table 2. Grid search result of Hierarchical model

• 3.1.4 Bayesian Optimization

Bayesian Optimization is an approach to tune hyper-parameters of a machine learning model which is considered a black box. It is an effective way to find the global optimum by using surrogate optimisation. The method tries to fit the truth model by sampling multiple points to create a posteriori distribution of a function (Gaussian process), a substitution model that best characterises the function to be optimised. And we can get a relative optimum of the substitution model. And then sample more points of the model iteratively to improve the posterior distribution. The algorithm is increasingly certain about which portions of the parameter space are worth examining and which regions are not as the number of observations rises, approaching the global optimum finally.

In HDBSCAN hyper-parameter optimisation, there are two parameters that need to be considered. In this case, the time complexity of grid search is $O(n^2)$, which is much bigger than Bayesian Optimization. So, Bayesian Optimization is a more sagacious way to optimise multiple hyper-parameters.

• 3.1.5 HDBSCAN (Bayesian Optimization)

In HDBSCAN, four Bayesian optimisers are set to approach the best v-measure score in four datasets, and the boundaries of the parameter are 2 to 35 for min_cluster_size and 2 to 10 for min_samples. There are a few differences in different datasets. The result is shown in Table 3. It can be observed that the numbers of noises are both extremely high in PCA features.

HDNSCAN	min_cluster_size	min_samples	V-measure Score	Noises
PGE_PCA	24	2	0.3034	3809
PGE_UMAP	25	4	0.5644	702
Resnet50_PCA	35	3	0.4295	3468
Resnet50_UMAP	24	10	0.6738	571

Table 3. Bayesian optimisation of HBDSCAN

4. Comparing and Evaluating Clustering Algorithms

The problem of model selection for clustering has been a challenge. One of the essential reasons is that there is no objective standard for clustering classification. Unlike classification problems, it has no reliable labels, while the definition of good or bad clustering results is often subjective. This also makes it difficult to compare different algorithms' results and complete the model selection for the clustering problem.

In this section, we will evaluate and compare the clustering quality of K-means, hierarchical clustering and HDBSCAN based on quantitative metrics and qualitative analysis separately.

4.1 Comparing and evaluating clustering quality by quantitative metrics

• 4.1.1 Evaluating the clustering quality based on ground truth

Some evaluation metrics can be used to assess the clustering results when the labels and ground truth are given. ARI and v-measure are typical representatives of such algorithms. In the following, we will evaluate the algorithm performance of k-means, hierarchical clustering and HDBSCAN with the ARI and v-measure scores data obtained from the model implementation, as shown in Figure 3, and all the data results are obtained when the hyperparameters have been optimised.

PGE PCA Feature:

Metrics	Kmeans	agg	HDBS
Silhouette	0.110918	0.084966	-0.173837
V-measure	0.446687	0.450976	0.303420
ARI Score	0.247427	0.252545	0.055845

Resnet50 PCA Feature:

Metrics	Kmeans	agg	HDBS
silhouette	0.137333	0.11110	-0.051614
V-measure	0.553885	0.57600	0.429458
ARI Score	0.344181	0.41253	0.103845

PGE UMAP Feature:

Metrics	Kmeans	agg	HDBS
Silhouette	0.552311	0.534228	0.409040
V-measure	0.567786	0.566360	0.564377
ARI Score	0.387077	0.393037	0.380434

Resnet50 UMAP Feature:

Metrics	Kmeans	agg	HDBS
silhouette	0.584901	0.556792	0.420960
V-measure	0.705007	0.729171	0.673765
ARI Score	0.565556	0.596148	0.478709

Figure 3. Clustering results

First, analysing the aspect of different deep neural network-based representations, the clustering results based on the Resnet50 dataset generally score higher than their counterparts in PGE when both the dimensionality reduction methods and clustering methods are the same. Then we can find that for different dimensionality reduction methods, the scores of UMAP are also generally higher than those of PCA methods, which indicates that UMAP retains the key features of data samples better.

Finally, with the same dataset and dimensionality reduction method (that is, under the same feature space), the clustering v-measure and ARI scores of the Hierarchical clustering are slightly higher than the one of K-means, and the scores of HDBSCAN are all lower than the first two, especially in PCA features.

This is probably because, unlike the first two clustering models, HDBSCAN will label some sample data points as noise, which are outliers, and return a labelled result of -1. However, most clustering evaluation metrics do not consider the situation of dealing with noise or outlier samples, which makes the HDBSCAN performance score suffer when using quantitative evaluation metrics. One

possible approach is to remove the relevant noise point data from the clustering results of HDBSCAN and use the evaluation metrics only for the clustered results themselves, thus deriving a relatively reasonable result, as shown in Figure 4. We can see that the scores are much higher than before and the others two clustering methods.

```
HDBSCAN performance without noise:


PGE UMAP data:
ARI score without noise:  0.47666164855563303
V-measure score without noise 0.6387625652611019
Total noise: 702

PGE PCA data:
ARI score without noise:  0.6720783692950607
V-measure score without noise 0.7525677322049146
Total noise: 3809

Resnet50 UMAP data:
ARI score without noise:  0.582979132201039
V-measure score without noise 0.7427004288761383
Total noise: 560

Resnet50 PCA data:
ARI score without noise:  0.8195767408851614
V-measure score without noise 0.8679004650317856
Total noise: 3468
```

Figure 4. HDNSCAN performance without noise

• 4.1.2 Evaluating clustering quality without ground truth

The problem with the metrics we used above is that they usually require prior knowledge of the ground truth class or need to be manually assigned by a human annotator as in a supervised learning environment; however, in practice, such preconditions are often unlikely to be satisfied. Therefore, it is also necessary to consider clustering scoring metrics without ground truth, such as the silhouette coefficient method.

Combining the results in the figure, we can find that the silhouette scores are generally lower than those evaluated with other clustering metrics in various datasets and dimension reduction methods. It is probably because the silhouette score prefers to calculate and reflect the compactness of the clusters, which leads it to be less sensitive to irregular and complex cluster shapes. On the other hand, it can also be observed that the silhouette method scores significantly higher in K-means clustering. That may be because, in the K-means algorithm, each cluster is defined only by its centre, which assumes in advance that each cluster is convex. The silhouette coefficient is naturally higher for convex clusters than other clusters; in contrast, in HDBSCAN, the scores can even be negative values. The scores in PCA data are very low because when reducing the dimension, PCA may lose some variance in different directions, and the PCA data is not well distributed.

4.2 Comparing and evaluating the clustering quality by qualitative analysis

To perform a qualitative comparison and analysis of the cluster quality, we calculated and visualised the percentage of tissue types separately under different cluster configurations, as shown in Figure 4-7.

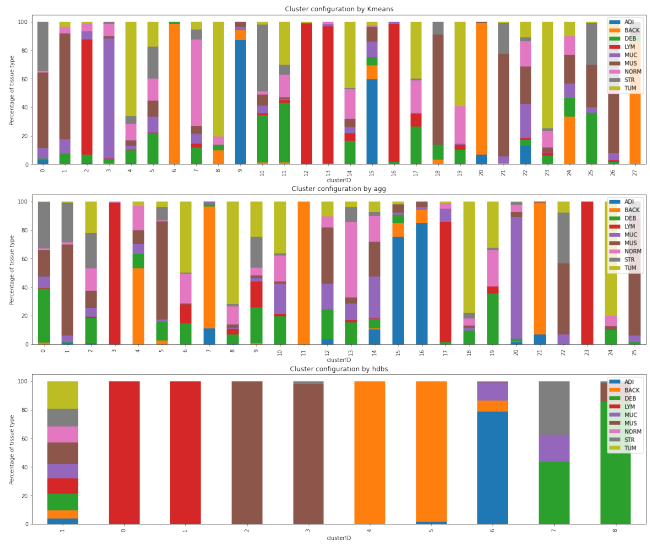


Figure 4. PGE-PCA cluster configuration

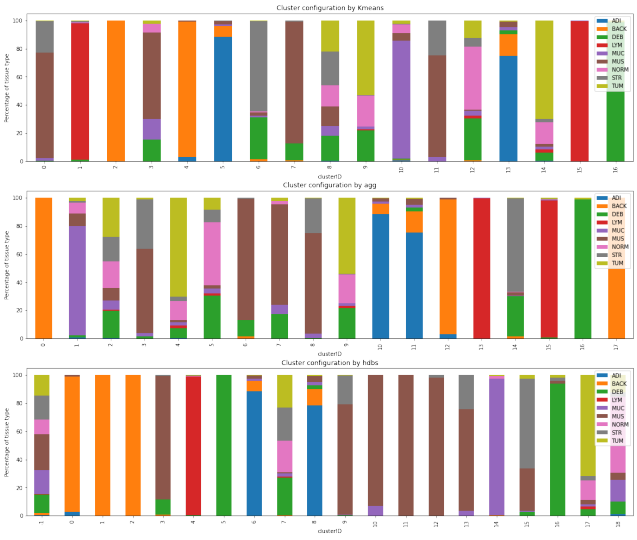


Figure 5. PGE-UMAP cluster configuration

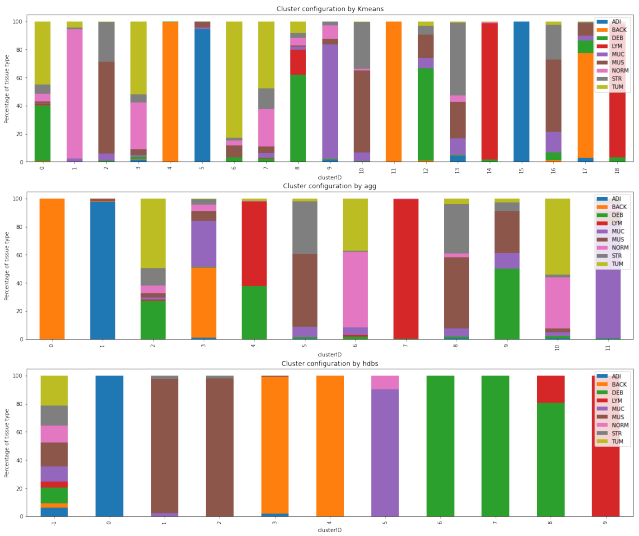


Figure 6. Resnet50-PCA cluster configuration

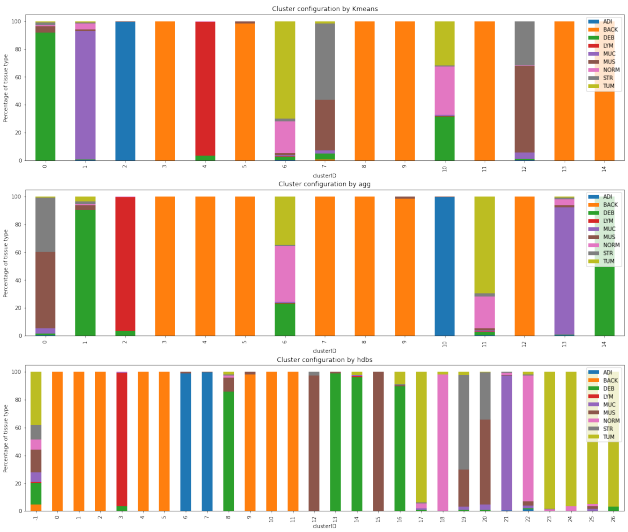


Figure 7. Resnet50-UMAP cluster configuration

The cluster class tissue abundance percentage level reflects a certain accuracy of clustering in an interpretable sense. The higher the cluster class accuracy, the more it can help us accomplish the initial exploratory filtering of a large number of tissue slice samples in colon carcinoma analysis and provide the basis for further data analysis.

These figures show that the cluster purity of the UMAP dataset is higher than that of the PCA dataset; the cluster purity of the Resnet50 dataset is also higher than that of the PGE dataset, and these validate our conclusions obtained through the quantitative metrics above.

In K-means and Hierarchical clustering, especially in the PGE dataset, it can be seen that there are different tissue classes mixed in the clusters, which indicates that the clustering results are not very good. The clustering algorithm may have incorrectly mixed the outliers into the clusters. While in HDBSCAN, the purity is relatively higher, and the clusters with label -1 are the noises recognised by the algorithm, which consists of different classes. It suggests that the noise points recognised by HDBSCAN are equally distributed in the feature space. If combined with the given ground truth label, we can conclude that samples labelled as outliers by HDBSCAN may appear on the image as a mixture of multiple tissue types with poor sample quality.

5. Conclusion

To summarise the analysis above, we can conclude that in PCA feature datasets, the K-means and Hierarchical clustering are better options, whereas the K-means model performs well in silhouette scores due to its algorithms; the Hierarchical clustering has a more reliable performance in v-measure and ARI scores. However, in UMAP feature datasets, HDBCAN has a better score than the others. On top of that, the scores without noises generated by HDBSCAN have a distinction performance. We believe that if there is a measurement to deal with the noises, for example, erasing the noises by calculating the distance with the noise and the centroid of clusters in HDBSCAN, the performance of this model will be much more outstanding.

6. Reference

[1]DBSCAN: Ester, M., Kriegel, H. P., Sander, J., & Xu, X. (1996, August). A density-based algorithm for discovering clusters in large spatial databases with noise. In kdd (Vol. 96, No. 34, pp. 226-231).

[2]HDBSCAN: Campello, R. J., Moulavi, D., & Sander, J. (2013, April). Density-based clustering based on hierarchical density estimates. In Pacific-Asia conference on knowledge discovery and data mining (pp. 160-172). Springer, Berlin, Heidelberg.