Machine Learning
Case Study - 1
Model Selection for Clustering

Agamjyot Singh Kohli - 2780992k
Anubhav Bhattacharya -
Priyadarshani Chimanchode - 2759607C
VIshrut Jain - 2781553J

Introduction

The following case study aimed to determine the appropriate model for Clustering. The process of drawing references from datasets of input data without labelled responses is known as unsupervised learning and Clustering is an unsupervised learning method. It is typically used as a method to identify the groups, generative qualities, and significant structures that are inherent in a set of instances.

The task of clustering is to divide the data points into several groups so that the data points within each group are more similar to one another and different from the data points within the other groups. It is essentially a grouping of items based on how similar and unlike they are to one another. Unsupervised learning is more susceptible to different clustering techniques and feature spaces than supervised learning is. Determining the quality of the cluster is similarly challenging.

In this exploratory work, we had to test several clustering algorithms with a variety of potential factors that may impact the number of clusters. On 5K colorectal patches represented by PathologyGAN, ResNet50, InceptionV3, and VGG16, we employed K-means and Lovian Clustering.

5,000 histological images of human colorectal cancer (CRC) and normal colon mucosa (NORM) as well as 9 tissue types are available for analysis. 4 feature sets, PathologyGAN, ResNet50, InceptionV3 and VGG16, are extracted to represent those 5,000 images' different dimensional feature spaces. To assess the quality of clustering solutions, several approaches are expected to be done and interpreted which include.

Methods

Two types of clustering algorithms are being used in this study.

1) K-means clustering
2) Louvain clustering

Four feature sets are extracted in the following task:

1. PathologyGAN
2. ResNet50
3. InceptionV3
4. VGG16

K-means Clustering

It is a type of unsupervised learning, which uses unlabeled data or data without defined categories or groups. The objective is to identify data in terms of groups. The K in k-means clustering indicates the number of groups that the algorithm needs to identify. Based on the features, the program iteratively assigns each data point to one of the K groups. Basically, the data points are grouped based on the similarity of their features.

Louvain Clustering

Louvain is a graph based algorithm that represents each data point as a node in the cluster, and each data point's similarity to another data point is represented by an edge.
The first step of the Louvain method is to place nodes in various clusters based on dense connections within a community known as modularity.
Then, until there is no Modularity improvement, clusters are repeatedly merged.

The number of clusters is influenced by a parameter in the modularity function called "resolution" (). This option may be changed to get different cluster counts.

Principle Component Analysisc(PCA):

A well-known unsupervised learning method for lowering the dimensionality of data is principal component analysis. While minimising information loss, it simultaneously improves interpretability. It makes data easier to plot in 2D and 3D and aids in identifying the dataset's most important properties.

Uniform Manifold Approximation and Projection (UMAP):

A fresh manifold learning method for dimension reduction is called UMAP (Uniform Manifold Approximation and Projection). The theoretical foundation for UMAP is based on Riemannian geometry and algebraic topology. The outcome is an efficient, scalable

algorithm that works with data from the actual world. With better run-time speed and visualisation quality that is comparable to t-SNE, the UMAP method may maintain more global structure. Additionally, UMAP may be used as a general-purpose dimension reduction strategy for machine learning because it has no computational limits on embedding dimension.
The main theoretical pillars of UMAP are topological data analysis and manifold theory. The language of topology and category theory adapts itself the most to the explanation of the theory.UMAP creates a topological representation of the high dimensional data by patching together its local fuzzy simplicial set representations and local manifold approximations. A comparable topological representation may be created using a similar procedure given any sort of low-dimensional data representation. In order to reduce the cross-entropy between the two topological representations, UMAP then optimizes the arrangement of the data representation in the low-dimensional space.

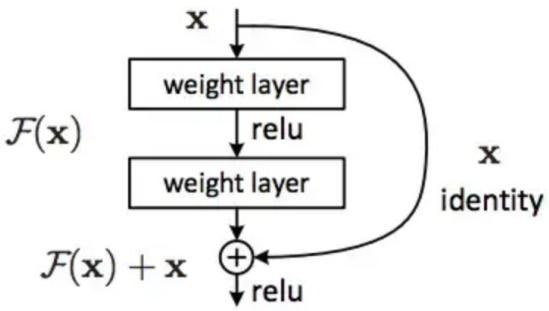
PathologyGAN

GAN or generative adversarial network is a technique that when given a dataset of images to train on can produce data with similar statistics as the training. For example, if the technique is given a set of photographs to train on, GAN can produce new images which are seemingly authentic to a human but are produced by the software.
PathologyGAN is a variant of GAN which augments the previously produced GAN variants to learn about cancer cell phenotypes through tissue architecture and cellular characteristics. PathologyGAN uses BigGAN as the base (BigGAN is proven to be widely successful in replicating datasets) but a few required changes were made to improve the Fr´echet Inception Distance (FID) and the architecture of the latent spaces, Elements of the styleGAN were also used to allow the generator to candidly develop the latent spaces and find the high-level features of the phenotypes.

Resnet50

Resnet50 is a convolutional neural network that is 50 layers deep. It is a variant of Resnet, which is a common neural network that is used for many tasks in computer vision. Resnet enables us to train very complex neural networks which are more than 150+ layers deep.

A major disadvantage with convolutional neural networks is the vanishing gradient problem, which comes into play when you increase the number of layers in the neural network. This problem is overcome by resnet by using ‘Skip Connection’.
Skip connection works by skipping over some layers of the model. This creates a residual block and resnets are formed by stacking these blocks together. Due to this, the output may differ.



Skip Connection

The idea behind skip connection is that if any layer hurts the performance of the model, then it can be skipped by regularisation.

Evaluation Techniques

In this case study, we have used two evaluation techniques:

1) V-measure
2) Silhouette

V-measure

This method was developed to evaluate the performance of a clustering algorithm which is usually a difficult task due to its nature. To use v-measure, we have to calculate two terms first :-

- Homogeneity - If each cluster has data points belonging to the same label, then it is called perfectly homogenous clustering. Homogeneity is the closeness of a cluster to being perfectly homogenous.
- Completeness - If all the data points belonging to the same class are categorized in the same cluster, then it is called perfectly complete clustering. Completeness measure how close our clustering is to being perfectly complete.

V-measure can be calculated as :

$$V_\beta = \frac{(1+\beta)hc}{\beta h + c}$$

Silhouette

This method calculates the average silhouette index for each sample, the average index for each cluster, and the overall average index for the entire dataset.

The silhouette value gauges an object's cohesiveness with its own cluster (cohesion) in comparison to other clusters (separation).
Silhouette Coefficient is defined as –

$$S(i) = (b(i) - a(i)) / (\max \{ (a(i), b(i)) \}$$

Where,

- a(i) is the average dissimilarity of i^{th} object to all other objects in the same cluster
- b(i) is the average dissimilarity of i^{th} object with all objects in the closest cluster.

Testing Strategy

All K-means models were test with K ranging from 3 to 25. Both V-measure and Silhouette scores were noted for all the models, and the models with maximum accuracy were selected in terms of both the metrics.

All Louvain models were tested with resolution ranging from 0.8 to 2. Both metrics were tested on each model and the models with best scores were selected as the best ones.

This technique ensured that we get the best K value for the k-means models and the best resolution for Louvain models.

Results:

Each representation was been reduced to two-dimensional 100 feature sets using both PCA and UMAP.
Each combination was then tested using the same two metrics, silhouette and v-measure.

The accuracy of each combination was calculated as follows:

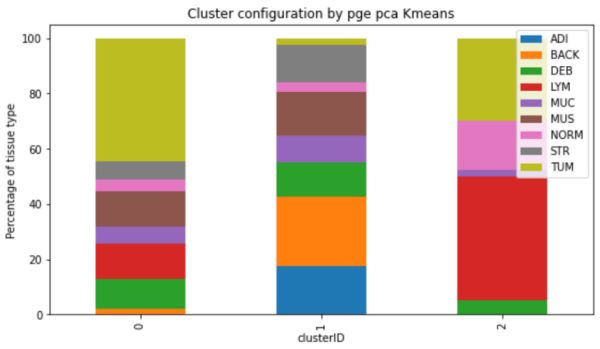
Best model scores for silhouette

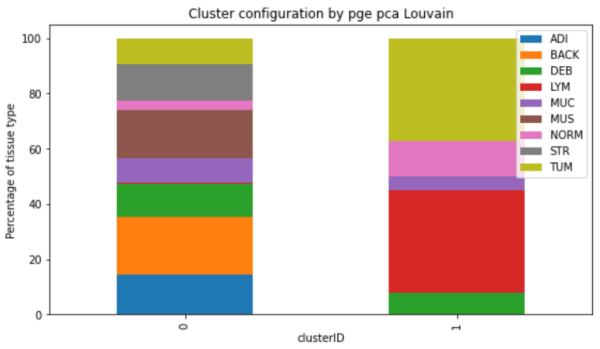
Metrics	pge pca Kmeans	pge pca Louvain	resnet pca Kmeans	resnet pca Louvain	pge umap Kmeans	pge umap Louvain	resnet umap Kmeans	resnet umap Louvain
silhouette	0.192233	0.303122	0.121973	0.116516	0.584298	0.385276	0.61517	0.359059

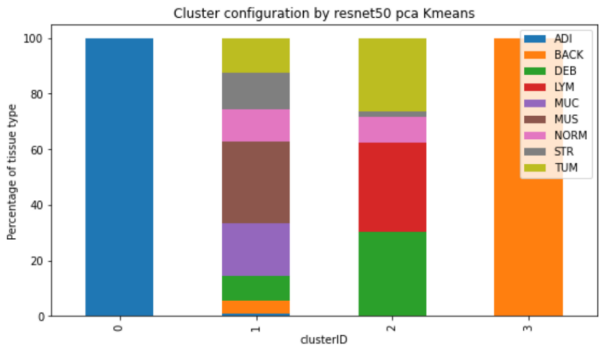
Best model scores for V-measure

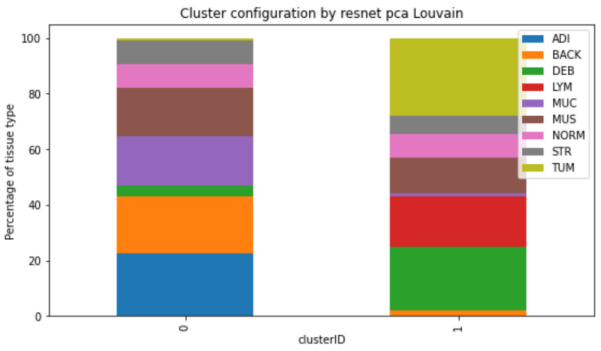
Metrics	pge pca Kmeans	pge pca Louvain	resnet pca Kmeans	resnet pca Louvain	pge umap Kmeans	pge umap Louvain	resnet umap Kmeans	resnet umap Louvain
V-measure	0.537212	0.582106	0.5946	0.649572	0.615517	0.644452	0.750753	0.686803

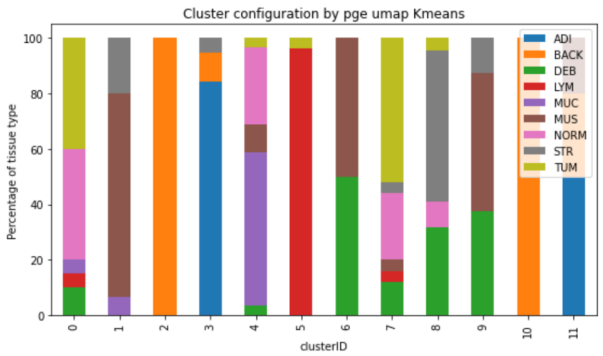
Following are the bar graph visualizations of each combination. They show the number of clusters formed by each algorithm and the grouping of each type of cells as well.

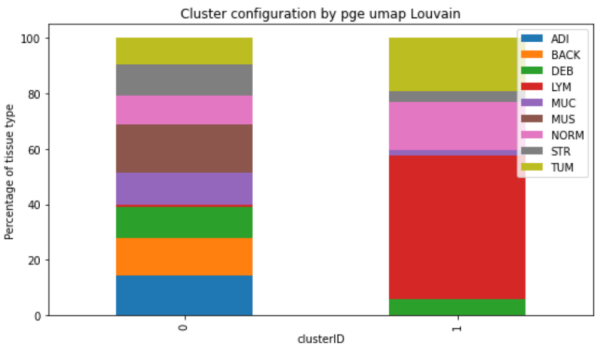


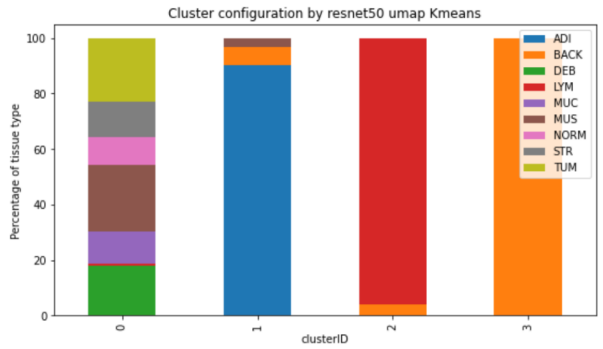


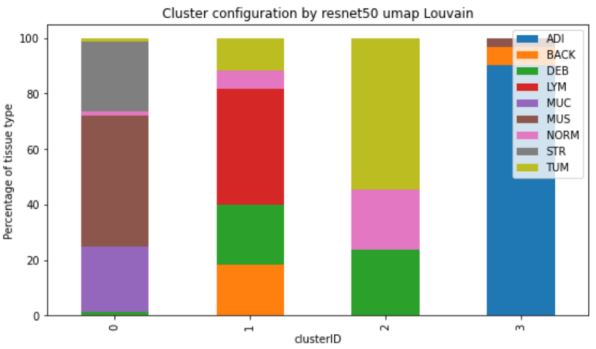












Conclusion:

UMAP and PCA algorithms were used for feature reduction. While UMAP was slower in performing the task, UMAP is still more efficient as It works incredibly well for visualizing clusters or groups of data points and their proximity to one another.

Feature extraction was done using PathologyGAN and Resnet50.
In terms of silhouette scores, PGE outperforms Resnet50 in almost all the cases.

The two clustering Algorithms used in the task were KMeans and Louvian but after observing the result of each case, it cannot be concluded which one is better for the model, as in some cases KMeans outperformed Louvain and in other cases vice versa.