Swarm Intelligence Clustering with CNN Feature Vectors for Breast Cancer Classification

Gokul Ganesan

1. Methodology

1.1 Dataset Acquisition and Preparation

The Breast Cancer Histopathological Image Dataset (BreakHis) was chosen for this project due to its comprehensive collection of breast cancer histopathology images, including both benign and malignant tumors. The dataset contains a total of 7,909 images across four different magnification levels: 40x, 100x, 200x, and 400x. The magnification levels are ignored for this analysis. Instead, we focus on the diagnostic subtypes of each breast cancer image. The images are classified into 2 groups, with 4 subgroups each. There are the benign and malignant groups, with each have 4 different types of tumors/cancers. We use these subgroups as our classification target. The goal is to use clustering to create an unsupervised learning model for classification of these different types of breast cancers, with high efficacy.

1.2 Feature Extraction with VGG16

The VGG16 Convolutional Neural Network (CNN) was chosen for feature extraction due to its established performance in image classification tasks. A simple CNN is better suited to this task as a more complex CNN trained on specific data may not pick up features from this data which has very low similarity to general training data. The network was initialized with pre-trained weights from ImageNet, and the final classification layers were removed to allow extraction of features from the global average pooling layer. This modification provided a 512-dimensional feature vector for each image, capturing high-level patterns and characteristics.

After normalizing and resizing the images, the VGG16 model processed each image to extract these feature vectors. The extracted feature vectors were used as input for subsequent clustering analysis. To add another dimension to the analysis, the VGG16 model was also fine-tuned to the breast cancer data by freezing the layers in the base model and training only the Dense layer we added. A dropout is included as well in order to ensure we do not overfit the model. This fine-tuning can help pick up features of images in this dataset specifically, since VGG16 has not been trained on classes similar to breast cancer. This can aid the clustering algorithm since the features will be more aligned/related to the important features of the images for classification.

1.3 Clustering Analysis

To classify the breast cancer images based on their extracted feature vectors, the KMeans clustering algorithm was initially employed. Before clustering, the feature vectors were standardized using StandardScaler to ensure that each feature contributed equally to the clustering process. Principal Component Analysis (PCA) was then applied to reduce the dimensionality of the feature vectors to 50 components, eliminating noise while retaining essential information.

To analyze the efficacy of the biologically inspired clustering algorithm, KMeans clustering performance provides a baseline to compare.

1.4 Swarm Intelligence

The algorithm we used for clustering is based on the behavior of flocking birds (like boids). One such algorithm, "Clusterflock" uses the way birds of different species flock together, as inspiration for clustering datapoints. This algorithm is mostly used in clustering for gene expression, or document similarity. It's use in this case to cluster based on CNN feature vectors is novel.

2. Results

2.1 Base VGG16

The base model started at a validation accuracy of just over 50%.

The KMeans clustering, with K=2, on the base VGG16 model, achieved a clustering accuracy of 50%.

	precision	recall	f1-score	support
benign	0.50	0.74	0.60	27145
malignant	0.51	0.27	0.35	27145
accuracy			0.50	54290
macro avg	0.50	0.50	0.47	54290
weighted avg	0.50	0.50	0.47	54290

This is essentially as effective as guessing, which was expected on the base model.

The KMeans model, with K=8 for the 8 cancer subtypes, had an accuracy of 18% which supports the idea that clustering on the features of the base model seem to be inconsequential, although marginally better than guessing. On these feature vectors, the swarm intelligence algorithm had too many distinct classes, since it acts similar to a KNN (no number of clusters).

2.2 Fine-tuned VGG16

The fine-tuned model performance was promising. It achieved a validation accuracy of 92% over 10 epochs on the classification of the two classes.

The swarm intelligence algorithm is meant to perform well in high dimensional data, and it seems like it does exactly that compared to KMeans. Here the KMeans algorithm maintains a 70% accuracy, after using the fine-tuned model, with K=2.

The Swarm Intelligence Model displays a much better ability to navigate high-dimensional spaces (Here, 50-D with PCA), with a 96.35% accuracy overall, outperforming the CNN alone.

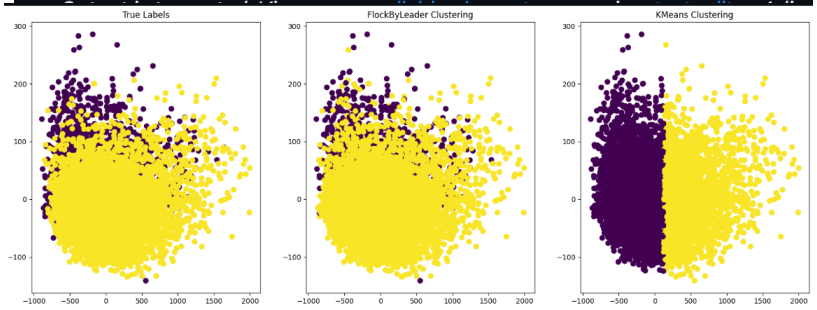
```
FlockByLeader Clustering Accuracy: 0.9635

FlockByLeader Classification Report:
              precision    recall  f1-score   support

      benign       1.00      0.88      0.94      2480
   malignant       0.95      1.00      0.97      5429

    accuracy                           0.96      7909
...

FlockByLeader Confusion Matrix:
[[2191  289]
 [   0 5429]]
```



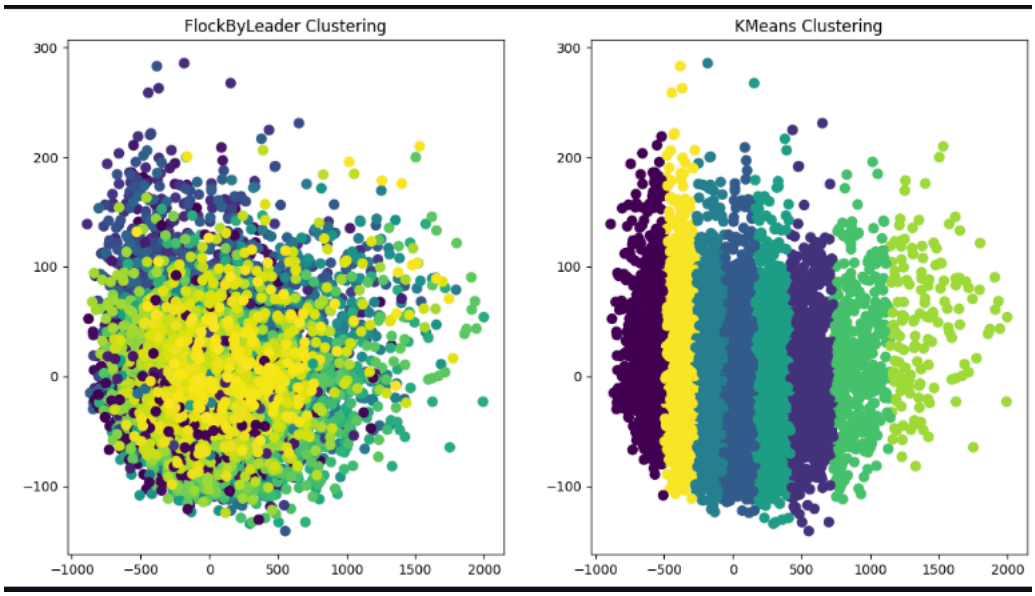
3. Analysis

The swarm intelligence algorithm performed extremely poorly on the base model's feature vectors because they would have been very sparse when computed on this dataset which VGG16 was not trained on. Therefore, the sparse datapoints lead to low similarities, and therefore too many distinct classes to be accurate. This is reinforced by the fact that the KMeans, which is forced to have 2 classes, had an accuracy of 50%, supporting the fact that it could not find clusters in the sparse data. But given the feature vectors of the fine-tuned model, it was more effective at identifying relationships (which is why it is so effective at identifying similar documents with tf-idf vectors).

This is exemplified when looking at the classification of points from both algorithms for the 8 subtypes of cancer. The CNN was not fine-tuned on each subtype, but instead on just the malignant/benign subgroups, which makes it interesting to see if those features alone can be clustered. The KMeans algorithm performs at 23% accuracy overall. The Swarm Intelligence Model, however, performs at 44% accuracy over all 8 classes.

4. Conclusion

This understanding of the swarm intelligence model can allow for more ideal classification models in the right contexts. Similarity measures like this, in high dimensional spaces can be optimized much farther with the swarm intelligence model than with traditional models like KMeans. One main reason for this is because the "ClusterFlock" Algorithm allows for non-linear clusters (similar to clustering algorithms like DBSCAN), which is where algorithms like KMeans fall short. This is better understood when looking at the predicted labels on the 2-component PCA chart for the 8 cancer subtypes (below). This chart exemplifies the non-linear aspect of swarm intelligence clustering, and the drawbacks of KMeans. With that, I have outlined the efficacy of swarm intelligence clustering in high dimensional contexts like CNN feature vectors.



Sources

"Clusterflock: a flocking algorithm", https://gigascience.biomedcentral.com/articles/10.1186/s13742-016-0152-3

"Flock by Leader: A Novel Machine Learning Biologically Inspired Clustering Algorithm", https://link.springer.com/chapter/10.1007/978-3-642-31020-1_15

"BreakHis." Kaggle, https://www.kaggle.com/datasets/ambarish/breakhis.

"Combining pretrained CNN feature extractors to enhance clustering of complex natural images", https://arxiv.org/pdf/2101.02767.