**Comparison of various image clustering algorithm in detection of leukemia**

**Aim of the Project:**

The aim of this project is to compare the statistics (like execution time, success rate) of different image clustering algorithms, such as K-median algorithm, against K-means algorithm. Our aim is to pave the way for future researchers to use this data to decide on which algorithm is most convenient for them. This also allows us, the researchers, into exploring the various algorithms involved in image-detection of leukemia and get to know more about them.

**Software Requirements:**

* MATLAB

**Purpose/Scope of this work:**

The main purpose of this work is to check the efficiency/effectiveness of different image clustering algorithms and compare their statistics side-by-side. This work can also act as reference/foundation for future work on other projects.

**Details of the Image dataset:**

* <https://www.kaggle.com/nikhilsharma00/leukemia-dataset>
* <https://homes.di.unimi.it/scotti/all/>
* <https://wiki.cancerimagingarchive.net/display/Public/SN-AM+Dataset%3A+White+Blood+cancer+dataset+of+B-ALL+and+MM+for+stain+normalization>

**Algorithm-1(Existing work):**

Assumptions: Microscopic images of leukemia positive and negative images are kept in different databases.

Input: RGB Blood sample smear containing WBC (white blood cells) and other cells. Unknown status of leukemia

Output: Leukemia positive/negative depending on the sample

**Step 1: Image Acquisition**

Input a microscopic image of a blood smear. RGB images are acquired from online databases.

**Step 2: Image Pre-processing**

Various pre-processing techniques are applied to the image to reduce the noise, adjusting the contrast and enhance the necessary features required for the segmentation. The RGB image is then converted to HSI model to enhance different colors in the image.

**Step 3: Segmentation**

Apply k-means clustering algorithm to the image to extract the different features of the image.

The steps involved are:

* Specify number of clusters *K* (taken as 4 in this case)
* Initialize centroids by first shuffling the dataset and then randomly selecting *K*data points for the centroids without replacement.
* Keep iterating until there is no change to the centroids. i.e assignment of data points to clusters isn’t changing.
* Compute the sum of the squared distance between data points and all centroids.
* Assign each data point to the closest cluster (centroid).
* Compute the centroids for the clusters by taking the average of the all data points that belong to each cluster.
* The approach k-means follows to solve the problem is called **Expectation-Maximization**. The E-step is assigning the data points to the closest cluster. The M-step is computing the centroid of each cluster.

It can be simplified to the formula:

Where wik=1 for xi if it belongs to cluster *k;* otherwise wik = 0. Also, µk is the centroid of xi’s cluster.

**Step 4: Feature Extraction**

The important features, that are necessary for the detection, are selected. The features extracted in the mentioned project are Energy, Correlation, Sum Entropy, Difference entropy, entropy, etc.

**Step 5: Classification**

Using the extracted features, a “classifier”, that has been trained to detect symptoms of leukemia, is then fed the data, which gives us the output based on its known classes.

**Limitations /Drawbacks Identified:**

1. The number of clusters have to be manually selected and may not be optimal for all cases.
2. The algorithm used (k-means) cannot be confirmed as the best approach to this problem, and thus, it has a chance of being an inefficient method.

**Sample Algorithm-2 (Proposed Work):**

**Step 1: Image Acquisition**

Input a microscopic image of a blood smear. RGB images are acquired from online databases.

**Step 2: Image Pre-processing**

Various pre-processing techniques are applied to the image to reduce the noise, adjusting the contrast and enhance the necessary features required for the segmentation techniques for each algorithm. The images may be converted to binary, grayscale, HSV and other such image formats to compare for optimal detection.

**Step 3: Segmentation**

Apply the algorithms to be used to the images and compute the execution results (execution times and success rates). Some methods that might be applied are:

1. **K-median algorithm:**

Given a set of points, the k-medians algorithm attempts to create k disjoint cluster that minimize the following equation. This means that the center of each cluster center minimizes this objective function.

This minimization is defined by the geometric median

1. **Mean-shift algorithm:**

It is a non-parametric clustering technique which does not require prior knowledge of the number of clusters.

Where K = Kernel,

n = number of data points,

h = Kernel bandwidth parameter and

d = dimension of space

**Step 4: Feature Extraction**

The necessary features are extracted and fed to a processing software (a classifier) to detect the presence of leukemia in the sample. Followed by details like execution time. The results from multiple sample images will then be compared to check for success rates and efficiency.

**Team members:**

|  |  |  |
| --- | --- | --- |
| **S.No** | **Reg.no** | **Name** |
| 1 | 19BCE1186 | Debojyoti Dey |
| 2 | 19BCE1202 | Avihrik Basak |
| 3 | 19BCE1758 | Pranish Shrestha |

**References:**

**For example:**

1. Prasidhi G. Fal Desai, Geeta Shet, (*Department of Electronics and Telecommunication, Goa University,(India))* “DETECTION OF LEUKEMIA USING IMAGE PROCESSING”, *International Journal of Advance Research in Science and Engineering,* Vol.07, Special Issue No.03, April 2018, pp. 149-156
2. V. Shankar, M. M. Deshpande, N. Chaitra and S. Aditi, "Automatic detection of acute lymphoblasitc leukemia using image processing," 2016 IEEE International Conference on Advances in Computer Applications (ICACA), 2016, pp. 186-189
3. Dr. R. Janaki, “Detection of leukemia in microscopic white blood cell images using gaussian feature convolutional visual recognition algorithm”, Journal of Critical Reviews, Vol. 07, Issue 3, January 2020, pp 173-180
4. DBScan and Gaussian mixtures with Gene Expression data for Acute Myeloid Leukemia, [*https://www.leukemiaairesearch.com/blog/detection-early-detection/article/dbscan-gaussian-mixtures-gene-expression-acute-myeloid-leukemia*](https://www.leukemiaairesearch.com/blog/detection-early-detection/article/dbscan-gaussian-mixtures-gene-expression-acute-myeloid-leukemia)
5. “The 5 Clustering Algorithms Data Scientists Need to Know”, Towards Data Science, *https://towardsdatascience.com/the-5-clustering-algorithms-data-scientists-need-to-know-a36d136ef68*