A K Means Based Genetic Algorithm for Data Clustering
Clara Pizzuti, Nicola Procopio

Bidhan Bashyal, Aarush Mathur

November 2, 2021

Overview

Overview

K Means Clustering

- Unsupervised data analysis technique in categorizing data
- K-means:most widely used clustering algorithms
- K-means involves following major steps:
  - 'n' number of centroid is selected (equal to the number of clusters
  - Objects are assigned to the cluster
  - Centroid is updated and 2 and 3 are repeated until no objects changes its membership to a cluster
- Number of clusters should be preassigned

Overview

Evolutionary K Means Clustering

- Genetic K means algorithm, Krishna K, Murty MN, 1996
- Capable of exploring search space and escaping from global minima
- Prior knowledge of number of clusters is not required
- High values of the evaluation indexes

Overview

CluGA(Propsed Method)

- Introduction
- Encoding
- Initialization
- Crossover
- Mutation
- Fitness Function
- Results

Introduction

- Consists GA integrated with K Means principle
- Each individual is a string of length equal to number of dataset objects $[2, \cdots k_{max}]$
- Crossover operator obtains the offspring by adopting a group based strategy
- Mutation operator reassigns an object x belonging to a cluster Ci to another cluster Cj if x is closer to Cj.

Encoding

- Uses Label based encoding
- Each object is in a position i of the chromosome and also associated with an integer in the alphabet[2, $\cdots k_{max}$]
- Renumbering procedure is used after application of genetic operator.
- Improves efficiency and avoids the presence of different string representing same solution

Initialization

- Uses approach used by $KMeans_{++}$
- $KMeans_{++}$ has different way of selecting center compared to regular K-Means
- $KMeans_{++}$ has following steps:
  - Choose the first cluster center$c_1 = x$ uniformly at random from the dataset X
  - Choose other cluster from remaining data objects
  - Repeat 2 until the k_max cluster has been found
- Has less convergence time, high speed and better accuracy

Overview

Project Goal

- Build a binary classifier to detect melanoma in images of skin lesions using deep CNN models
- Use images within the same patient
- Use patient-level contextual information (metadata)
- AUC evaluation metric

THANK YOU