INTRODUCTION

Cluster Analysis:- Cluster is a group of objects that belongs to the same class. In other words, similar objects are grouped in one cluster and dissimilar objects are grouped in another cluster.

Goal-

Analyse the performance of different clustering algorithms on SMS collection dataset where each SMS is categorized either HAM or SPAM

Dataset-

All the analysis has been performed on SMS Spam Collection from UCI machine repository. There are total 5574 messages.Memory Size of Dataset=500kb. Data is classified as Spam and Ham. Stopwords are also considered while performing the analysis. Baisc Statistics of the dataset can be seen in **Table 1**

Total Message	5574
SPAM	4827
HAM	747
DISTINCT TERMS	13098

Table1:- Basic Statistics of Dataset

Method-

Messages were classified as ham and spam. Further the messages were tokenized and stopwords were also considered since it gives impact on the accuracy of the algorithm. Ham is labelled as 0 and Spam is labelled as 1. Sequence Matcher algorithm is used for evaluating the similarity matrix. Following algorithms are used for clustering the data into Spam,Ham.

- **K-Means**

  Initially, two cluster points were selected and the similarity score for both the points with all the dataset points was calculated using SequenceMatcher algorithm. Further the points with minimum similarity score was selected and was assigned to the cluster. Initial Cluster points are user defined.

- **DBSCAN**

  Intially a random point is selected and then points which have similarity less than the particular eps(The maximum distance between two samples for them to be considered as in the same neighborhood) are selected in neighbourhood points. The minimum number of points considered in a neighbourhood point were selected as core point.The input given over here for DBSCAN algorithm is the similarity matrix between the selected point and the dataset.

- **Agglomerative Clustering**

  The euclidean distance is calculated between the similarity matrix. No. of clusters are selected as two. Affinity Measure is selected as Euclidean also linkage is considered average for evaluating the corresponding labels for all the instances.

- **BIRCH**

  BIRCH clusters a point without having to check against all other data points or clusters. It also remove outliers and produce good clusters with a single scan of dataset. There are three types of affinity measures in BIRCH which can majorly impact on the accuracy of the model. It is also assumed that it is beter than other algorithms for large datasets.

- **Spectral Clustring**

  Data points are repesenteds as the vertices V of a graph G. Vertices are connected by edges E . Edges have weights W. Large weights mean that the adjacent vertices are very similar, small weights imply dissimilarity.

Parameter Tuning

Parameter Tuning of different clustering algorithms is done i.e, parameters are given different value and corresponding accuracy and time complexity is evaluated for each value. Eventually the parameters with value which gives the best accuracy compared to the other values is selected as the result for that clustering algorithm. Parameter Tuning for each different clustering algorithms can be clearly visualised from **Table 2**

	Tol	**Iterations**	**Accuracy**	**Time**
KMeans	0.0089	1800	43.30%	255.44sec
	0.000005	900	56.70%	41.23sec
	0.000000117	2700	56.00%	492.31sec
	eps	**min_samples**	**Accuracy**	**Time**
DBSCAN	0.05	10	62.90%	0.090sec
	0.045694	10	86.60%	30.5sec
	0.0004	15	13.50%	43.33sec
	Threshold	**Branching Factor**	**Accuracy**	**Time**
BIRCH	0.0005	80	41.00%	29.32sec
	0.00005	70	65.91%	39.32sec
	0.000005	60	70.23%	39.78sec
	Linkage	**Affinity**	**Accuracy**	**Time**
Agglomerative Clustering	Average	Euclidean	86.60%	448.21sec
	Ward	manhattan	86.60%	36.6sec
	complete	cosine	30.01%	55.37sec
Spectral Clustering	**Eigen_Solver**	**Affinity**	**Accuracy**	**Time**
	None	rbf	56.27%	137.77sec
	arpack	nearest_neighbor	43.12%	12.21sec
	arpack	rbf	56.26%	512.34sec

Table 2:- Parameter Tuning for different clustering algorithms

Results

DBSCAN and Agglomerative clustering gives the same accuracy which is better than other algorithms but DBSCAN proved to be the best algorithm for my dataset as it gives the highest accuracy with comparitively less time complexity than other algorithms whereas Spectral Clustering gives very low accuracy when compared to other algorithms which can be clearly seen in **Table 3**.

	Parameter1	**Parameter2**	**Accuracy**	**Time**
Kmeans	0.000005	900	56.70%	41.23sec
DBSCAN	0.045694	10	86.60%	30.5sec
BIRCH	0.000005	60	70.23%	39.78sec
Agglomerative Clustering	Ward	manhattan	86.60%	36.6sec
Spectral Clustering	None	rbf	56.27%	137.77sec

Table 3:- Representation of different clustering algorithms with their corresponding optimal parameters value

Conclusion & Discussion

Agglomerative Clustering seems to be the most efficient algorithm for Spam Filtering as it further evaluates the affinity between the similarity matrix which helps it to cluster it further correctly, DBSCAN is said to be efficient for large datset and hence gives the high accuracy with less time complexity. KMeans, Birch, Spectral Clustering are inefficient as Kmeans cannot maintain the accuracy throughout the dataset.

The accuracy can be further improved if, other algorithms are used for evaluating the similarity matrix for the message dataset. Also other different clustering algorithms can be used to increase the accuracy and reduce the time complexity for the efficient clustering of the messages.