Comparison of F1-Score of Clustering Algorithms on IRIS Classification

Sumit Mishra
AITS ML Intern
AI-Technology & Systems
www.ai-techsystems.com
sumit.mishra0432@gmail.com

***Abstract*— Iris dataset is one of the basic datasets. It contains data of various species of flower of Iris plant. SepalLength, SepalWidth, PetalLength, PetalWitdh and Species are the data contained in this data set. It includes three iris species that are Iris-Setosa, Iris-Versicolor, Iris-Virginica with 50 samples each as well as some properties about each flower. One flower species is linearly separable from the other two, but the other two are not linearly separable from each other. It's the Assignment-3 given to me as a Machine Learning Intern. I have done the Exploratory data Analysis, Preprocessing, modelled three clustering algorithms and have compared the f1 scores and performances of these 3 Clustering Algorithms.**

Keywords—Clustering, Mean Shift, Agglomerative Clustering, K-means, iris, Confusion matrix.

I. INTRODUCTION

Clustering is a technique to categorize the data into groups. Distance metrics plays a very important role in the clustering process. There are number of algorithms which are available for clustering. Clustering algorithms are generally not preferred for the simplest reason being they fall under unsupervised learning. Such unsupervised algorithms do not perform well on less data. We are clustering algorithms for supporting the same hypothesis of unsupervised learning. We use 3 clustering algorithms namely K means, Agglomerative and Mean Shift and produced the f1 scores for each. We then compare the f1 scores of the Algorithms and plotted a bar graph for the comparison.

The Data set and EDA section consists of information about the dataset and some exploratory data analysis using the matplot library and seaborn library. Feature Engineering section consists of feature engineering. Splitting of dataset section consists of splitting of datasets and building model and training on the data. Section V consists of comparison study of the different models.

II. DATASET & EDA

A. Dataset

It contains data of various species of flower of Iris plant. SepalLength, SepalWidth, PetalLength, PetalWitdh and Species are the data contained in this data set. It includes three iris species that are Iris-Setosa, Iris-Versicolor, Iris-Virginica with 50 samples each as well as some properties about each flower.

Four features were measured from each sample: the length and the width of the sepals and petals, in centimeters. Based on the combination of these four features, the clustering algorithms are need to be designed.

B. Exploratory Data Analysis

We plotted a scatter plot which shows the relationship between the petal width and petal length.

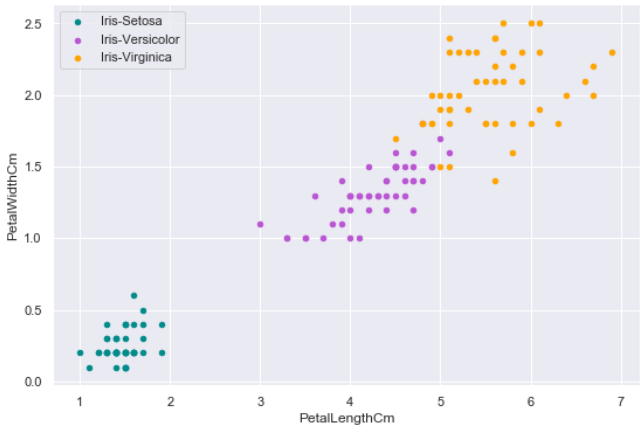

Fig 1.0

In the above figure we see that the Iris Setosa is linearly separable from other two classes.

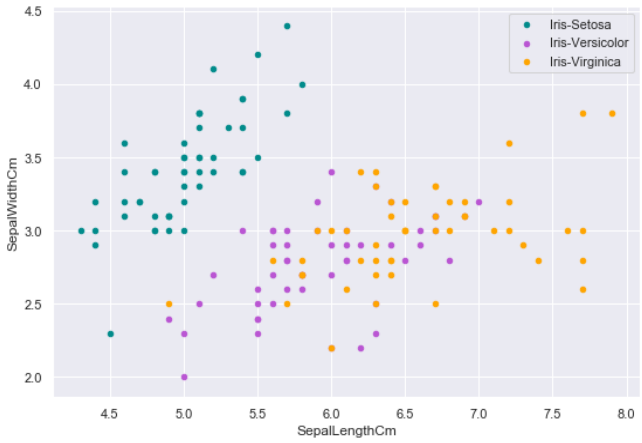

Fig 2.0

So, we conclude that the iris-Setosa is the most distinct flower and a cluster can easily be formed around the Setosa. Hence any clustering algorithms might be able to get a good prediction about the Iris-Setosa than the other two.

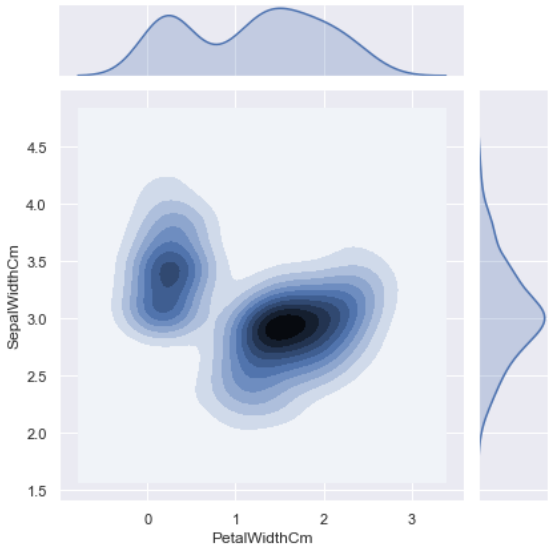

Fig 3.0

Fig 3.0 shows the joint plots that intuitively shows the cross relation between the features i.e. Sepal width and petal width.

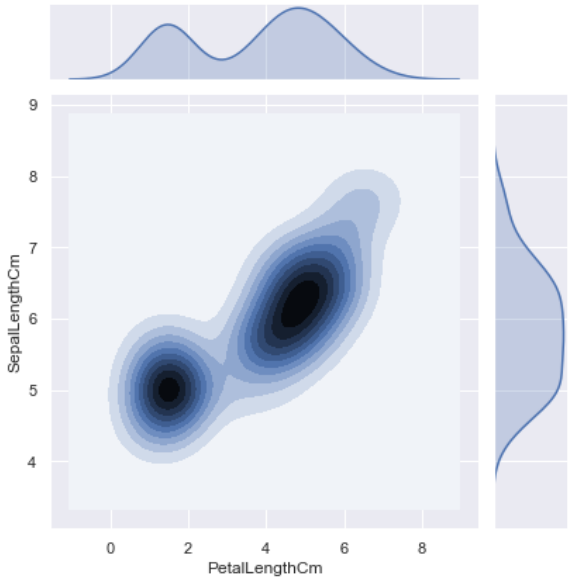

Fig 4.0

Fig 4.0 shows the joint plots that intuitively shows the cross relation between the features i.e. Sepal length and petal length.

III. Feature Engineering

Iris Data Set has150 examples and just 4 features to train our model. When we'll trained the model on just 4 features which are sepal length and width, petal length and width. I have used the Boosted gradient descent algorithms to get the intuition about the feature importance and plotted the feature importance.

From fig 5.0, We came to know that petal length is the most important feature which is followed by petal width. Hence the Petal length will have the most impact on the Clustering Algorithms.

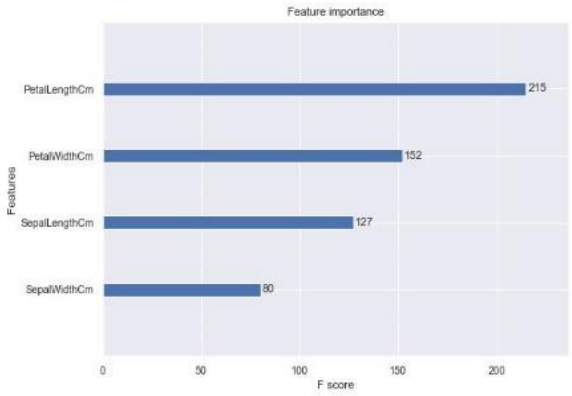

Fig 5.0

IV. Splitting Dataset & Model Building

We split the dataset where 70% is chosen for training and 30% for testing. So, out of 150 data examples, we have 105for training and 45 for testing. Hence shape of training data is (105,4) & (45,) for test. After splitting, we choose our clustering algorithms and build the models and fit on the training data. The hyper parameters for each model are shown below.

KMeans (n_clusters = 3, init = 'k-means++', max_iter = 300, random_state = 0)

k-means is a method of vector quantization, originally from signal processing, that is popular for cluster analysis in data mining. *k*-means clustering aims to partition *n* observations into *k* clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the clusters. This algorithm aims at minimizing an objective function know as squared error function. I have used a loop to iterate over the number of clusters and hence taken the best model for the final prediction and from the fig 6.0 three clusters in the algorithm is best performing.

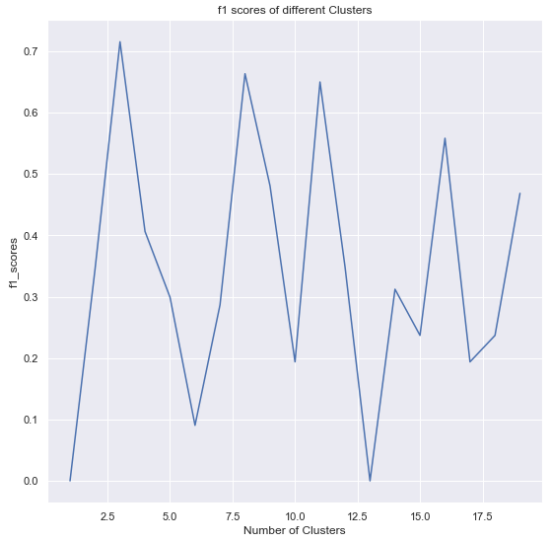

Fig 6.0

Mean shift Clustering Algorithm sims to discover "blobs" in a smooth density of the samples. It is a centroid-based algorithm, which works by updating candidates for centroids to be the mean of the points within a given region. The hyper parameters used for mean shift during training are.

MeanShift (bandwidth = bandwidth, bin_seeding = True)

In statistics and data mining, Agglomerative Clustering is a hierarchal clustering algorithm. It recursively merges the pair of clusters that minimally increases a given linkage. The hyper parameters used during training are.

AgglomerativeClustering (n_clusters = 3, affinity = 'euclidean', linkage = 'ward')

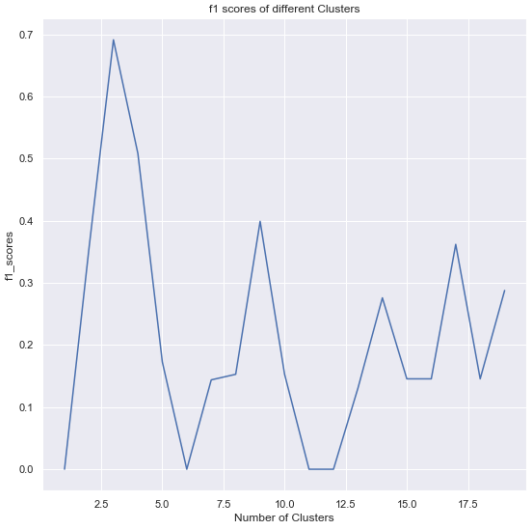

Fig 7.0

I have used a loop to iterate over the number of clusters and hence taken the best model for the final prediction and from the fig 6.0 three clusters in the algorithm is best performing and from the fig 7.0 it can be seen that at the number of clustering of 3 is the best performing.

V. COMPARISION OF F1 SCORES

After training the Clustering Algorithms, we do prediction on the test dataset and we obtain the following classification reports on each model.

	precision	recall	f1-score	support
0	0.00	0.00	0.00	0
1	1.00	1.00	1.00	16
2	0.82	1.00	0.90	18
3	0.00	0.00	0.00	11
accuracy			0.76	45
macro avg	0.45	0.50	0.47	45
weighted avg	0.68	0.76	0.72	45

Fig 8.0 F1 Score of KMeans

	precision	recall	f1-score	support
0	0.00	0.00	0.00	0
1	1.00	1.00	1.00	16
2	1.00	0.67	0.80	18
3	0.00	0.00	0.00	11
accuracy			0.62	45
macro avg	0.50	0.42	0.45	45
weighted avg	0.76	0.62	0.68	45

Fig 9.0 F1 Score of Mean Shift

	precision	recall	f1-score	support
0	0.00	0.00	0.00	0
1	1.00	1.00	1.00	16
2	1.00	0.72	0.84	18
3	0.00	0.00	0.00	11
accuracy			0.64	45
macro avg	0.50	0.43	0.46	45
weighted avg	0.76	0.64	0.69	45

Fig 10.0 F1 Score of Agglomerative

I have plotted a bar graph to compare the f1 scores of the Clustering Algorithms.

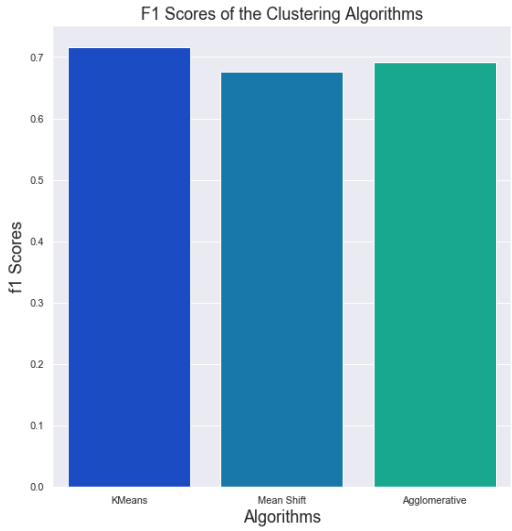

Fig 11.0 Bar graph of the f1 scores.

VI. CONCLUSION & FUTURE SCOPE

We conclude that the kmeans clustering algorithm is the best performing algorithm among the three of the given. Clustering Algorithm are not a good approach when the data is too less as it will not be able to make clusters with a good accuracy. For iris, Classification algorithms are the preferred option over the Clustering Algorithms. When the amount of data is large then the Clustering based algorithms can work pretty well.

A. Author and Affiliations

AITS affiliation (www.ai-techsystems.com)

REFERENCES

[1] Performance Metrics"https://towardsdatascience.com/data-science-performance-metrics-for-everyone-4d68f4859eef?gi=aec260b0d6af"

[2] Wikipedia "https://en.wikipedia.org/wiki/K-means_clustering".

[3] Kaggle.com "https://www.kaggle.com/ashishs0ni/iris-dataset"

[4] Sklearn "https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html"

[5] IEEE conference paper format "*https://www.ieee.org/conferences/publishing/templates.html*"