Comparison of F1 Scores and Performance of 3 Clustering Algorithms on Iris Data Set

Sumit Mishra

AI Tech Systems

7 August, 2019

New Delhi, India

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 Specis 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.**

General terms :

Algorithms, Measurement, Performance

Keywords :

Clustering, Accuracy, KNN, EDA

I. INTRODUCTION

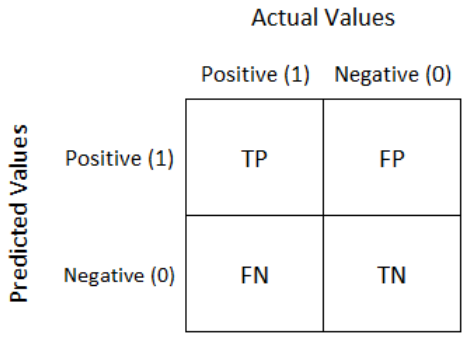
Clustering is a technique to categorize the data into groups. Distance metrics plays a very important role in the clustering process. There are number of algorithms which are available for clustering. In general, K-means is a heuristic algorithm that partitions a data set into K clusters by minimizing the sum of squared distance in each cluster. The algorithm consists of three main steps: a) initialization by setting center points (or initial centroids) with a given K, b) Dividing all data points into K clusters based on K current centroids, and c) updating K centroids based on newly formed clusters. It is clear that the algorithm always converges after several iterations of repeating steps b) and c).

II. METRICS

- the F_1 score (also F-score or F-measure) is a measure of a test's accuracy. It considers both the precision and the recall of the test to compute the score:

$$F_1 = 2 * \frac{precision * recall}{precision + recall}$$

- A confusion matrix is a table that is often used to describe how a model has performed on the given dataset for which true values are known. It is also known as error matrix and it allows the visualization of the performance of an algorithm. It gives us insight not only into the errors being made by a classifier but more importantly the types of errors that are being made. The number of correct and incorrect predictions are summarized with count values and broken down by each class.



- Prevalence - A machine learning algorithm with very high sensitivity and specificity may not be useful in practice when prevalence is close to either 0 or 1.

III. FIGURES AND TABLES

- This scatter plot shows the relationship between the petal width and petal length. By this we can say that the Iris Setosa is linearly separable from other two classes.

Figure -3.1 Scatter Plot b/w petal width and length

- This scatter plot shows the relationship between the sepal width and sepal length. As seen above the Iris Setosa is linearly separable when petal length and width are plotted but the scatter plot between the Sepal length and Sepal Width are not linearly separable as shown below.

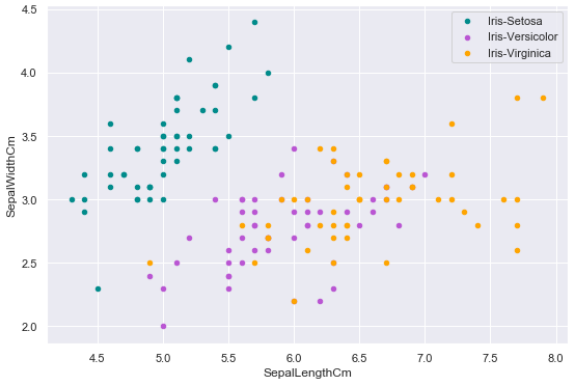


Figure -3.2 Scatter plot b/w sepal width and length

- The Heatmap of the correlation matrix has also been plot to know the feature dependency. By this we know if there's a positive or negative dependency of the features.

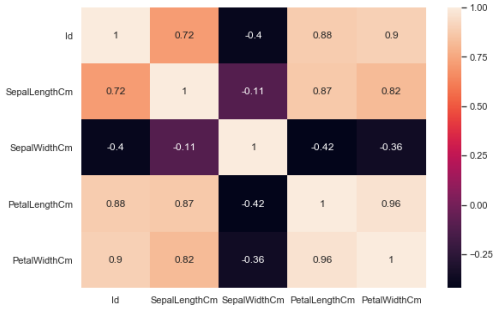
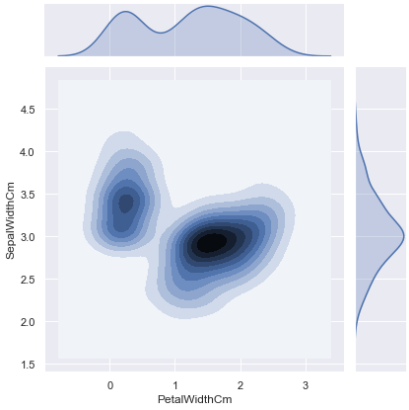


Figure -3.3 Correlation Matrix Heatmap

- There are also the joint plots that shows the cross relation between the features i.e. Sepal length and petal length, Sepal width and petal width.



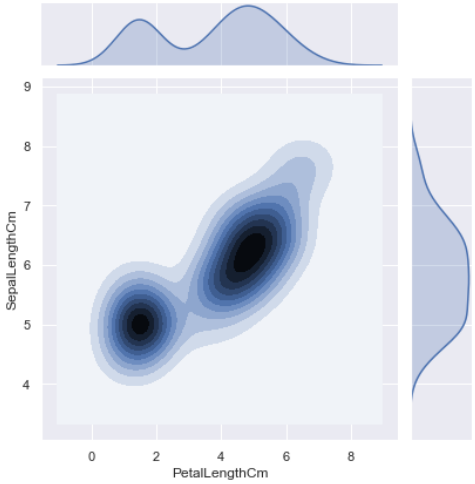


Figure 3.4 Kernel Density Plots 2

IV. CLUSTERING TECHNIQUES

3 Clustering models are used which are K-Means Clustering, Mean Shift Clustering and Agglomerative Clustering.

4.1 K-Means Clustering

It 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 given by.

$$J(V) = \sum_{i=1}^{c} \sum_{j=1}^{c_i} \left(\|x_i - v_j\| \right)^2$$

4.2 Mean Shift Clustering

It aims to discover “blobs” in a smooth density of 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.

$$m(x) = \frac{\sum_{x_i \in N(x)} K(x_i - x) x_i}{\sum_{x_i \in N(x)} K(x_i - x)}$$

4.3 Agglomerative Clustering

This algorithm is a hierarchal clustering algorithm. It recursively merges the pair of clusters that minimally increases a given linkage distance.

V. RESULTS AND EXPERIMENTS

- I have used the Boosted Gradient Descent to get insight of the feature importance and plotted the feature importance bar graph. From the graph is can be clearly seen that the Petal Length is the most important feature in the dataset.

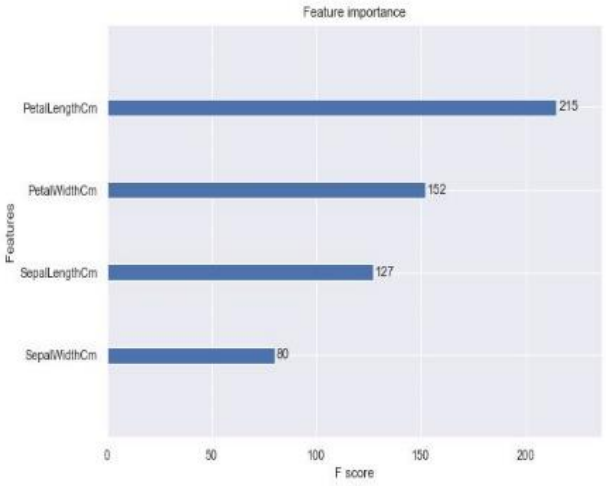


Figure -5.1 Feature Importance

- K-means Clustering - I have used a loop to iterate over the number of clusters and hence taken the best

model for the final prediction.

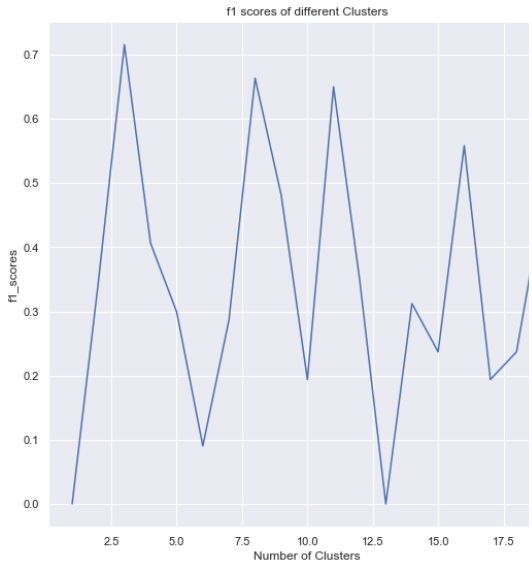


Figure-5.2 K-mean Clustering Performance

K-Means Algorithms gives good f1 score of 0.7156 and an accuracy of 75.56%.

- Mean Shift Clustering -The mean shifting Clustering Algorithms has an accuracy of 62.22% and a f1 Score of 0.6756.
- Agglomerative Clustering- I have also used a loop to iterate over the number of clusters and used the best model for the final prediction.

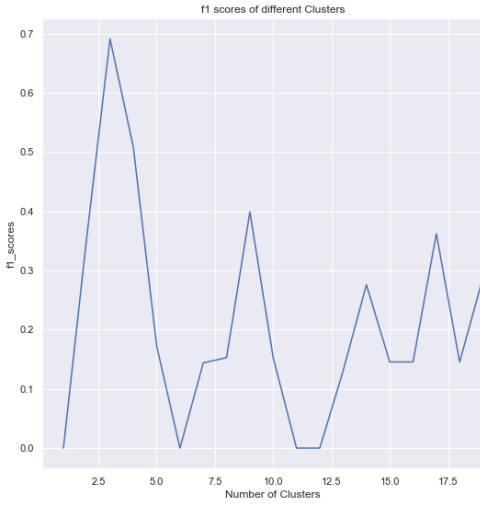


Figure -5.3 Agglomerative Clustering performance

Agglomerative Clustering had an accuracy of 64.44% and a f1 Score of 0.6910.

I have plotted a bar graph that compares the f1 scores of the three clustering algorithms.

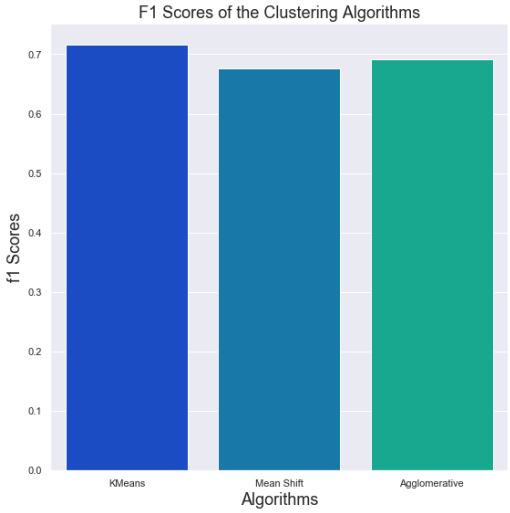


Figure-5.4 f1 score comparison

The K-Means Clustering Algorithm is the best performing model with a f1 score of 0.7156.

VI. AUTHOR AND AFFILIATION

Artificial Intelligence Tech Systems affiliation (www.ai-techsystems.com)

VII. CONCLUSION

I have performed the Exploratory Data Analysis on the Iris Data set and Preprocessed it for the modelling and successfully Modelled 3 Clustering Algorithms and compared their f1 Scores and Found out that the K-means is the best performing clustering algorithm on the iris data set with a f1 Score of 0.7156.

VIII. ACKNOWLEDGEMENT

I would like to express my deepest appreciation to all those who provided me the possibility to complete this

report. A special gratitude I give to AITS for giving me this opportunity. This project also enhanced my presentation skills.

IX. REFERENCES

1. Performance Metric for everyone(https://towardsdatascience.com/data-science-performance-metrics-for-everyone-4d68f4859eef)
2. IEEE conference paper format (*https://www.ieee.org/conferences/publishing/templates.html*)
3. Kaggle Iris Data Set (*https://www.kaggle.com/ashishs0ni/iris-dataset*)
4. AI Tech Systems (*www.ai-techsystems.com*)