Target Variables in Clustering

Jayden Yap Jean Hng

School of Computing

Singapore Polytechnic

Singapore

JAYDENYAP.21@ichat.sp.edu.sg

Abstract

Earth is a huge place and is host to thousands if not millions of species of flora. With that many species it was a challenge for Mankind to be able to group flora apart and a big part of Research was finding differences between species of flora. Today I will be finding a way to most optimally cluster flora, finding out whether leaving the target variable in the dataset is good for clustering or not. Target variable is a vague term, many use it in other machine learning problems like Classification or Regression. However, Target variable in clustering terms can refer to the variable that already clusters variables and is the main way to cluster them. Today we will find a way to cluster these using or without using that variable

I. INTRODUCTION

We will make use of Unsupervised Machine Learning to improve predictions of Iris flower species types. The type of Unsupervised Machine Learning that I will use is known as Clustering. This involves fitting the training data onto the model such as KMeans or Birch Clustering and then generating the cluster labels, then one could use the labels to generate visualisastions of the clusters or the centroids, centroids are the centers of each cluster Therefore, we need to build cluster models.

II. DEFINITIONS

A. Machine Learning

Machine Learning or automatic learning is a scientific field, and more specifically a subcategory of artificial intelligence. It consists of letting algorithms discover “patterns”, namely recurring patterns, in data sets. This data can be numbers, words,images or statistics. Types of Machine Learning include: Supervised Learning, Unsupervised Learning and Reinforcement Learning.

B. Unupervised Learning

Unsupervised learning is associated with learning without supervision or without training. In unsuperised learning the algorithms are trained with data which is neither labelled nor classified. In unsupervised learning, the agent needs to learn from patterns without corresponding output values. Unsupervivsed learning can be either Clustering or Association.

III. RELATED WORKS

There have been several previous studies and discussions conducted on the issue of using target variables in clustering machine learning problems such as the one we are conducting today. One such discussion was conducted in 2018 by several authors at the Data Science exchange program website. (Tanguy, A,2018). They scrutinized various ML approaches for the development better algorithms for clustering when one’s dataset already has a target variable that can help with the clustering or be considered a heavier weight.

Another one was by Siti Khotijah in 2019. (Siti, Khotijah. (2019) In this paper commonly used data mining and machine learning techniques and their complexities are summarized.Another study was conducted as well in this, where they used the clustering algorithm known as K Means clustering for a dataset very similar to the one I am going to investigate. It also investigates species of living beings on Earth as a dataset.

IV. METHODLOGY

A. Abbreviations and Acronyms

ML may be used to abbreviate Machine Learning

Similarly, abbreviations for some model names such as KNN for K Nearest Neighbors, SBM for Support Vector Machines may be used. These are widely used in the industry.More abbreviations such as K-Means or PCA for Principal Component Analysis may be used.

B. Units

We will be using SI units in our investigation unless otherwise stated such as with our data's unit measurements such as mg/dl for cholesterol values. We will also be using Silhouette score later for our model evaluation. This metric has no units.

C. Equations

The silhouette score of a cluster's labels such as K Means or other cluster algorithms used in the industry can be calculated as s(i) with the following rules:

$$s(i) = \begin{cases} 1 - \dfrac{a(i)}{b(i)} & \text{if } a(i) < b(i) \\ 0 & \text{if } a(i) = b(i) \\ \dfrac{b(i)}{a(i)} - 1 & \text{if } a(i) > b(i) \end{cases}$$

$$s(\boldsymbol{o}) = \frac{b(\boldsymbol{o}) - a(\boldsymbol{o})}{\max\{a(\boldsymbol{o}), b(\boldsymbol{o})\}}$$

Figure 1: Silhouette Score metric equations

IV. DATA SOURCE

The dataset I will use for this investigation is from the Machine Learning repository UCI, a prestigious website commonly used by Data Scientists to obtain datasets and learn more about Machine Learning. https://archive.ics.uci.edu/ml/datasets/iris The dataset

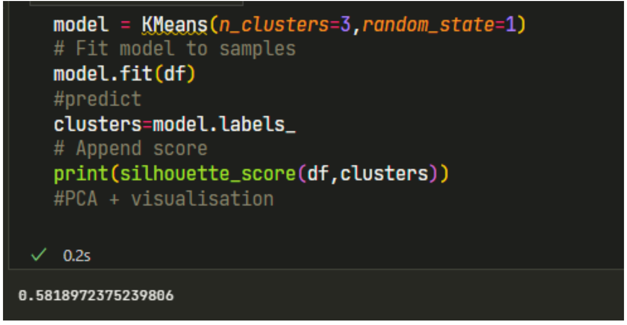
is provided by R..A Fisher ,, and the donor is Micheal Marshall. This dataset is very old and originated from 1936. .It has no missing values and I intend to perform unsupervised machine learning on it with the method of Clustering

V. DISCUSSION

My investigation started by haing to choose whih Programming Language to execute it in. One popular

language used by Data Scientists these days is the R language. However, me as the author was more familiar with the use of a language called Python, also very popular for data scientists and is actually one of the most popular programming languages in the world. I decided to perform the execution of the task with Jupyter notebook which is a file type that allows for cleaner presentation of code as well as markdown to truly have a more professional and beautiful and elegant look to the code. So I made use of Python packages like sklearn for machine learning algorithms and pandas for dataframe management. I continued by importing the data into the notebook. Fortunately this is one of the many many datasets that SKLearn provides for free on the package. I can import it by using load_iris package. After that I implemented the use of models like K-Means clustering, a popular algorithm used by many data scientists alike today.

From Figure 2, we can see that we obtained a silhouette score of 0.58 which is a very good result for our K Means algorithm clustering.



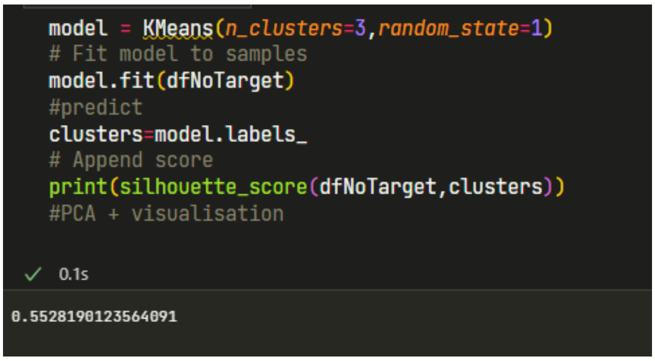
```
model = KMeans(n_clusters=3,random_state=1)
# Fit model to samples
model.fit(df)
#predict
clusters=model.labels_
# Append score
print(silhouette_score(df,clusters))
#PCA + visualisation

✓ 0.2s

0.5818972375239806
```

Figure 2: KMeans with Target Variable

Now we need to try fitting this data without the target variable



```
model = KMeans(n_clusters=3,random_state=1)
# Fit model to samples
model.fit(dfNoTarget)
#predict
clusters=model.labels_
# Append score
print(silhouette_score(dfNoTarget,clusters))
#PCA + visualisation

✓ 0.1s

0.5528190123564091
```

Figure 3: KMeans without target variable

We can see that our silhouette score metric calculated with the formula previously stated in part II was resulting in a lower amount of score. The difference can be calculated with the formula X-Y=difference where X is the first result from figure 2 and Y is the result we get from Figure 3. The difference equates to about 0.03 which is a significant decrease in performance. This shows that the target variable species helped a significant amount in clustering as it resulted in a better clustering algorithm.

VI. CONCLUSION

We can conclude from this investigation that using the target variable in datasets such as the Iris dataset used in this inventisgation is helpful for clustering as it helped to increase the score.

REFERENCES

[1] Tanguy., Khan, Annony Mousse, Kasra Manshaei (2018). Is it possible to cluster data according to a target variable in Unsupervised Machine learning?

[2] https://datascience.stackexchange.com/questions/30860/is-it-possible-to-cluster-data-according-to-a-target Seh, Adil Hussain. (2019). A Review on Heart Disease

https://datascience.stackexchange.com/questions/30860/is-it-possible-to-cluster-data-according-to-a-target

[2] Siti Khotijah (2020) Iris Dataset Clustering using the K-Means algorithm https://www.kaggle.com/code/khotijahs1/k-means-clustering-of-iris-dataset