**Codon Usage Classification by Machine Learning**

**Naren Hemanth  
n01426222**

**Introduction:**

A codon usage is the frequency with which each codon is used in each DNA sequence. This information can be useful for several applications, including understanding patterns of gene expression. By comparing the codon usage patterns of different species, researchers can identify changes in codon usage that may be associated with adaptive evolution. By the DNA sequence we can generate the number of times each codon occurs in the sequence. Codons are sequences of three nucleotides that correspond to a specific amino acid in a protein. There are 64 possible codons in the genetic code, but not all of them are used equally. By analyzing the codon usage patterns in each DNA sequence or set of sequences, researchers can gain insights into the underlying biological processes at work. The analysis by codon usage patterns in a DNA sequence is to predict the expression levels of specific genes. For example, if a particular codon is used more frequently than expected, it may indicate that the corresponding gene is being highly expressed.

**Dataset Description:**

This dataset consists of a set of sequences of the frequency of the codon based on different kingdoms. The number of times each codon occurs in the DNA sequence helps to calculate the relative frequency of each codon. The data not only consists of the DNA types but also consists of Kingdom category, Species ID, Number of Codons and Species Name. In this Project we are working in predicting the frequency of the codon with the Kingdom categories. Let us discuss about each and every column in the dataset.

Column 1: Kingdom: The 'Kingdom' is a 3-letter code of 11 CUTG database names.   
Column 2: DNA type: Denoted as an integer for the genomic composition of 13 species.  
Column 3: SpeciesID: Given as integer, which uniquely indicates the entries of an organism.   
Column 4: Ncodons: Number of occurrences with the species is the codon frequencies. Column 5: Species Name: Different species names.  
Column 6-69: codon frequencies ('codon') including 'UUU', 'UUA', 'UUG', 'CUU', etc., are recorded as floats (with decimals in 5 digits).

**Classification by Machine Learning:**

* **Supervised Algorithm -> Random Forest Classification**

Codon usage can vary between different organisms and can affect the efficiency and accuracy of expression levels of specific genes.One way to analyze codon usage is using Random Forest Classification, a machine learning method that make predictions. To use random forest classification for analyzing codon usage, you would first need to gather a dataset of codon usage. So, from this dataset we are considering the Kingdom category as the target with all the codon frequencies.

Next, we are checking with both Kingdom and the DNA type where through the bar plot, we can observe that the codon frequency has much pattern with the Kingdom compared to the DNA type. By the above bar plots we can analyze that the codons are more frequent with the kingdom.

Chart, bar chart

Description automatically generated Chart, bar chart

Description automatically generated

Next, we would need to pre-process the data, which involves cleaning and formatting the data, normalizing numerical features, and encoding categorical features. We would then split the dataset into training and testing sets. After working on the data, we finally got the accuracy of 0.90. Which can be considered as good accuracy. Now we shall run the same dataset in the Weka and choose the random classifier and the output as below.

Table

Description automatically generated

By using the Weka tool with the same dataset, we got the Correctly classified instances as 90% and the root mean squared error as 0.145. If we get the values of the root mean less than 1 then we can consider the accuracy level as very high. So, comparing both the Random Classifier and the Weka tool we can analyze that they both retrieve the same accuracy which defines as there is best pattern of codons with the Kingdom category of the UGCT database.

* **Unsupervised algorithm -> Kmeans Classification.**

K-means clustering can be used to group codons based on their usage patterns in different kingdoms. By using this algorithm, we shall group the codons into clusters to get the similar data points together. The performance of the K-means clustering algorithm will depend on the quality and size of the dataset, as to get the accurate level we can measure by considering different clusters or else by finding the centroid in the group codons based on their usage patterns. In the context of codon usage analysis, the elbow method can be used to identify the optimal number of clusters to use when grouping codons based on their usage patterns in different organisms.

Chart, line chart

Description automatically generatedChart, scatter chart

Description automatically generated

As we had divided the codon patterns into four clusters where we can find the centroids in the clusters as in the above image. By further calculating by taking n\_clusters = 10 and creating the labels we got the accuracy level as 0.94. Which is the best accuracy. If we compare the same dataset with the Weka tool KMeans analysis we get the clustered instances as 55% and 45%.

Graphical user interface, application

Description automatically generated

As by using the python in the Jupyter we can choose the number o clusters as of we want but while coming to the Weka application tool it has chosen only two clusters as shown. So, it has divided whole the codon patterns in to only two clusters and found the centroids. Even though by using the python we got the accuracy level of 94 by taking 10 clusters.

* **Principal Component Analysis – Feature Reduction Algorithm.**

By projecting the data onto a new set of dimensions, termed principal components, that represent the most significant fluctuations in the data, principal component analysis (PCA), a statistical approach, can be used to reduce the dimensionality of a dataset. PCA can be used in the context of codon usage analysis to pinpoint patterns and trends in the data and to depict the connections between various codons and their usage. As I have divided into two components namely Principal Component 1 and principal component 2. As by observing the previous data analysis I had chosen the main three targets as 5, 4 and 1 from the kingdom and to analysis the codon frequency. In the context of codon usage analysis, the explained variance ratio can be used to understand how much of the variation in codon usage is captured by each principal component. By using the explained variance ratio to interpret the results of the PCA analysis and to understand the patterns and trends in the data. As

Chart, scatter chart

Description automatically generated

As with the use of only few data information we have got the variance ration of 0.41 and 0.07. So, to get more variance we can add some other data and can run the same process of the PCA. Adding few data will help us to get the variance more compared to the first dataset. So, the new variance after adding the data to the dataset we got 0.56 and 0.21. which results as the codons are having high patterns with the 5,4,1 and 3 kingdom organisms.

**Conclusion:**

The aim of this study was to investigate the use of machine learning algorithms for predicting the frequency of codons in a codon usage dataset based on the kingdom categories. To achieve this goal, two different algorithms were applied to the data: random forest classification and K-means clustering.

The results of the analysis showed that both algorithms were able to achieve high levels of accuracy. Specifically, the random forest classification obtained a correctly classified instances rate of 90% and a root mean squared error (RMSE) of 0.145, while the K-means clustering algorithm obtained an accuracy level of 0.94 when the number of clusters was set to 10 and labels were created for the clusters.

However, it's important to note that the accuracy of a machine learning model will depend on a variety of factors, including the quality and size of the dataset, the appropriateness of the chosen model and features, and the appropriate hyperparameter tuning. Therefore, it's possible that the K-means clustering algorithm may not always outperform the random forest classification for thisdataset, and it may be necessary to try different approaches and techniques to find the best solution for your specific research question.

In addition, the results obtained using Weka (i.e., a clustered instances rate of 55% and 45%) may differ from the results obtained using other tools or programming languages. This could be due to differences in the implementation of the algorithms, the parameters used, or other factors.

Overall, these results suggest that both the random forest classification and K-means clustering algorithms were effective approaches for analyzing the codon usage dataset and predicting the frequency of codons based on the kingdom categories. The random forest classification obtained a correctly classified instances rate of 90% and a root mean squared error (RMSE) of 0.145, while the K-means clustering algorithm obtained an impressive accuracy level of 0.94 when the number of clusters was set to 10 and labels were created for the clusters. These results demonstrate the potential of machine learning algorithms for understanding codon usage patterns and trends in different organisms. Further research and analysis may be needed to explore the capabilities of these algorithms more fully and to identify the most effective methods for addressing different research questions

**--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------**