Solid State Nanopore Data Classification with Singular Value Decomposition

Chen Chen, MinGyu Kim

December 4, 2022

1 INTRODUCTION

Our goal through this project is to classify 5 different bio-structures (plasmid, dsDNA, dsRNA, ssRNA, and ribosome) using the singular value decomposition (SVD) method and explain why our method is a reliable model. Previously, researchers have used machine learning techniques like the random forest model, and they applied the data directly to the black-box classifiers. Thier approaches presented a high accuracy, however, they were not able to provide a proof of reliability of their models because of the limit of the neural networks' innate characteristics. To address and solve this issue, we propose to use the mathematical method, SVD, which estimates the eigenvectors of each feature matrix in distinct dimensions. With the eigenvectors, a linear regression model is introduced to calculate the error between the test data and estimated values. This approach will help us find a correct prediction by allowing us to know the error that differentiates each feature value of one bio-structure from feature values of the other bio-structures.

2 METHODS

2.1 Data

For this project, we use five classes biomolecular data from solid state nanopore, which are DNA plasmid puc19 (2284 events), dsDNA 1kb ladder (919 events), ssRNA ladder (1874 events), dsRNA ladder (1398 events), and ecoli ribosome (17077 events). We choose 17 features in order to do the classification, which are a1_pA, s1_pA, s3_pA, pot_sec, dwell_sec, t12_sec, mean_amp_pA, mean_amp_nS, max_amp_pA, max_amp_nS, med_amp_pA, std_amp_pA, area_pA_sec.

Additionally, in order to have the high quality events, we select the events with an SNR larger than 5. Finally, we have 2258 events for DNA plasmid puc19, 901 events for dsDNA 1kb ladder, 1865 events for ssRNA ladder, 1382 events for dsRNA ladder, and 16932 events for ecoli ribosome. We do not take buffer events into account in this project.

2.2 Singular Value Decomposition (SVD)

For each class, we firstly split the data into training (80%) and testing (20%) sets. Each class has a matrix X made by the training data, whose rows are features and columns are events. Then by following equation 1, we conduct SVD for each class separately, where U is a unitary matrix or eigenvectors that represent features, Σ is singular values that are sorted in the order of importance, V is a complex unitary matrix, m is the number of features, and n is the number of events.

$$X = U_{m\times m} * \Sigma_{m\times n} * V^T_{n\times n} \tag{1}$$

```
# python code :
u, sigma, v = numpy.linalg.svd(data sample X)
```

2.3 Residual Norm Error

The residual norm error is calculated by subtracting the estimated value from the true value.

$$L = min_{w_i}||z - \sum_{i=1}^{k} w_i * u_i|| = min_W||Z - U_k W|| \tag{2}$$

where z is a test data or input data, w is a weight vector, and U is the unitary matrix of the training data X. For each class, we find the estimated weight matrix W and separately compute the error with every test data input.

2.4 Least Square Regression Optimization

To obtain the estimated weight, we optimize a linear regression model that replaces the estimated term $U_k W$. In a linear regression model, we can express that

$$\bar{Y} = \sum_{i}^{m} w_i * u_i = W * U_k$$

where w is weight, u is the unitary matrix, and y is the estimated feature vector.
Least square error is

$$L = ||Y - \bar{Y}||^2 = ||Y - U_k W||^2 = (Y - U_k W)^T (Y - U_k W)$$

In order to get the minimum error, the gradient of the error function must be zero (local minima).

$$\frac{\partial L(W,U)}{\partial W} = -2U_k^T Y + 2U_k^T U_k W \overset{SET}{=} 0 \tag{3}$$

$$\hat{W} = (U_k^T U_k)^{-1} U_k^T Y \tag{4}$$

Using Equation (2),

$$L = ||Y - U_k \hat{W}|| = ||Y - U_k * (U_k^T U_k)^{-1} U_k^T Y|| \tag{5}$$

$U_k^T U_k$ is an orthogonal matrix, so this term can be regarded as an identity matrix. Therefore, the error term L from Equation (2) can be reduced as

$$L = ||Y - U_k U_k^T Y|| \overset{replace Y with Z}{=} ||Z - U_k U_k^T Z|| = ||(I - U_k U_k^T) Z|| \tag{6}$$

3 Results and discussion

We calculated the error with the test data "Z" from all classes using Equation 6, and the data with minimum errors were selected as our predictions.

3.1 Classification results with different k

3.2 Explainable SVD

In the following, we will discuss the eigenvectors for each class (i.e., matrix U) aiming to explain how does SVD work for the classification.

4 CONCLUSION

[1]

5 Division of Labor

References

[1] Marc-André Gonze et al. “Modelling the dynamics of ambient dose rates induced by radiocaesium in the Fukushima terrestrial environment”. In: *Journal of Environmental Radioactivity* 161 (2016), pp. 22–34. DOI: https://doi.org/10.1016/j.jenvrad.2015.06.003.