Music Genre Classification

Abhinav Khanna
Student Researcher
akhanna@princeton.edu

Amandeep Singh
Student Researcher
email@princeton.edu

Abstract

1 Introduction

2 Related Work

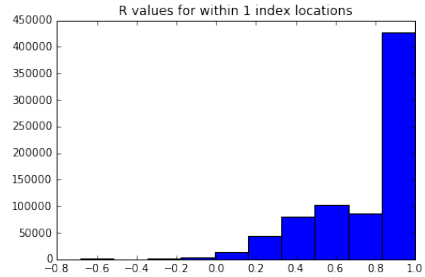
3 Methods

3.1 Description of data

The training data used in this paper comes from a set of 33 reference samples that have their entire genome bisulfite sequence. We primarily focus on the data for chromosome 1 for our analysis. The data consists of 33 reference samples with their methlyation proportions for each CpG site on the chromosome. We utilize this information to impute the values for the missing methlyation values in the test chromosome sample.

To better understand the relationships between different tissues, we computed the Pearson R correlation coefficient (Figure 2). Certain tissues have a meaningful, non-pure chance, correlation with other tissues suggesting that certain tissues may be more apt at predicting given test values than other tissues. We also know from Figure 1 that locations that are closer together have higher correlations, and are more similar to each other than locations that are further apart on the chromosome.

Figure 1: Pearson R Correlation for neighboring index locations



Examining the distribution of the mean methylation values across different tissues the distribution appears to have a right skew, and looks like it could be representative of a beta distribution. We believe that attempting to model the similarity between our test tissue and the training tissues as a problem of judging which prior beta distribution the test sample is most likely to have been drawn from may be a worthwhile approach. The mean methylation values can be seen in Figure 2. Furthermore, plotting the methylation values across a given tissue by location yields a graph that does not look linear, suggesting that a non-linear regressor may yield a better result (Figure 3).

Figure 2: Table of Correlation Values for Different Tissues

	0	1	2	3	4
0	(0, 0)	(0.712834867218, 0.0)	(0.757290076957, 0.0)	(0.708640674253, 0.0)	(0.377274968645, 0.0)
1	(0.712834867218, 0.0)	(0, 0)	(0.703314264614, 0.0)	(0.706602501721, 0.0)	(0.329787127457, 0.0)
2	(0.757290076957, 0.0)	(0.703314264614, 0.0)	(0, 0)	(0.738807452418, 0.0)	(0.351294390129, 0.0)
3	(0.708640674253, 0.0)	(0.706602501721, 0.0)	(0.738807452418, 0.0)	(0, 0)	(0.312640590975, 0.0)
4	(0.377274968645, 0.0)	(0.329787127457, 0.0)	(0.351294390129, 0.0)	(0.312640590975, 0.0)	(0, 0)

Figure 3: Mean Methylation across 33 Tissues

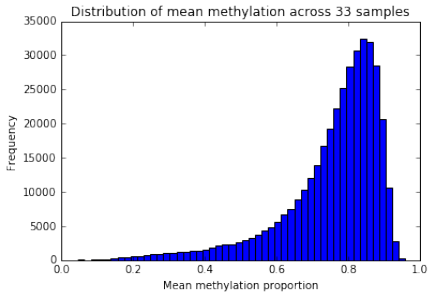
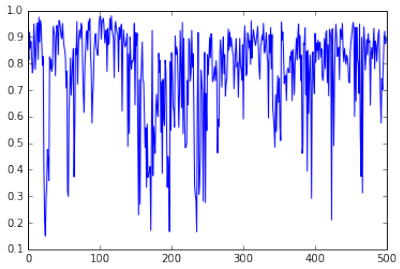


Figure 4: Methylation Values across a Single Tissue



3.2 Imputation Pipelines

The basic structure of our approach remained the same across pipelines. We began by calculating the similarity between the test tissue and the training tissues. The similarity was judged through some type of similarity metric, and different pipelines used different metrics. After measuring similarity, the pipelines combined the information from the most similar tissues to predict the values of the test tissue. The process of combining the information was also varied from pipeline to pipeline.

3.2.1 Similarity Metrics Tested

1. Beta Distribution Prior Probability
2. Pearson's R correlation

3.2.2 Combination Methods Tested

1. Weighted mean
2. Support Vector Regression
3. Decision Tree Regression
4. Random Forest

The pipelines that were run with the beta prior segmented the data into window size chunks, and used those chunks to train the regressors for the given location values.

3.3 Evalulation

The effectiveness of each pipeline was judged by examining the Pearson R correlation of the filled in test vector to the full vector, and computing the root mean squared error between the two vectors.

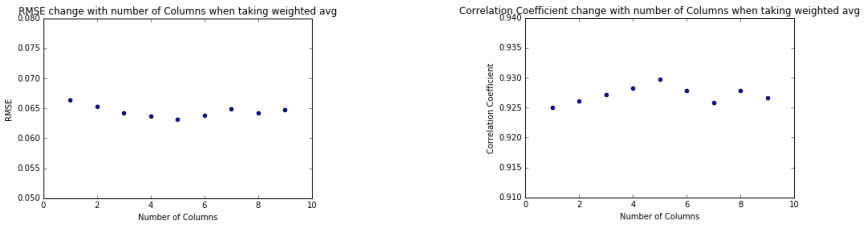
4 Results

We tested the pipelines against chromosome 1, and graded the accuracy of the imputation based on the RMSE, and Pearson's R correlation values. The following error rates were achieved by predicting the test sample for chromosome 1. Performance was variable with the weighted mean with correlation similarity performing the best out of all the pipelines.

Imputation Accuracy		
Pipeline	**RMSE %**	**Pearson's R**
BPS + SVR	42.6	32.2
BPS + DT	17.7	0.58
BPS + RF	17.7	0.58
CS + WM	6.3	0.93

For the weighted mean with correlation similarity pipeline, we optimized how many most similar vectors we looked at in order to impute the test value by graphing the change in the RMSE and the Pearson's R correlation with different top most similar vectors. As figure 5 showcases, the lowest RMSE and the highest correlation occur when 5 columns are considered. For the beta prior similarity pipelines, we segmented the data sets by a given window size, it was predicted that smaller window sizes would perform better because the data would appear more linear on smaller windows, and through trial and error, it was discovered that window size had no effect on the final accuracy.

Figure 5: RMSE and Pearson R change with the number of columns



5 Discussion and Conclusion

Acknowledgments

We would like to acknowledge Princeton University for offering the wonderful COS 424 class that provided us with the opportunity to take on this research. We would also like to thank our Professor, Professor Englehart, and her TA team for providing us with guidance and the tools necessary to make this project happen.

References