Midterm REPORT on Phenotypic Prediction from Transcriptomic Features

Team:

Jay Lohokare(111492930)

Nittin Aggarwal(111401512)

Revati Damle(111461639)

Vaibhav Mathur(111447903)

Summary:

The data-set shared with us consisted of output from Salmon, using which we had to build a model to correctly classify the data.
Before starting with creating the machine learning model, we created a data file with every row representing the data corresponding to one sample, having last column of the row as the class of population it belonged to.

By running SVM on the entire data-set, we found out that TPM and Effective length were much better parameters to classify the data-set compared to length. We also proved that TPM and Effective length were correlated features, hence we selected only TPM for the further steps in creating a better learning model.

Then, we used decision tree to find out the features that were the most dominant ones in classification of the data set. Decision tree gave us the **34** most contributing features (We have attached the average values of TPM, Effective length and length for all these 34 features at end of this report).

We then discarded all other features from the data set and ran SVM, Random forests, Decision tree classification models for classification. For SVM, we selected gamma value =1/34, and tested using RBF, linear and polynomial kernel(s).

To validate our results, we did a 5 fold cross-validation for all models that we used. To further validate the model, we split the data into 2 part (80% + 20%), trained the models using 80% data and then tested the model using the 20% data previously unknown to the model.

Without reducing the data, we got following results for 5 fold validation – Decision tree (Average 67%), SVM (Linear kernel – Average 61%), Random forests (Average 63%).

Using 5 fold validation on the reduced number of features, we got following results for classification: SVM (Average 79.9% for linear kernel), Decision tree (Average 82%), and Random forests (Average 73%)

This report includes the results of all classification models and k-fold cross validation.

Results

1. Decision Tree to reduce dimensions, then using SVM, RF, DT as classifiers.

Initial shape of dataset: (369, 199325)

Size of data after feature reduction using decision tree: (369, 34)

A. **Decision Tree 5 Fold Results:**
Scores: [0.79220779 0.85526316 0.86111111 0.77777778 0.83333333]

Mean: 0.823938634465

F1 Scores: [0.81956486 0.83837535 0.83079273 0.76363937 0.84381808]

F1 Mean: 0.819238078988

B. **SVM 5 Fold Results for kernel :**

Linear: Scores:[0.83116883 0.75   0.84722222 0.77777778 0.79166667]
Mean: 0.799567099567

F1 Scores:[0.83083862 0.75444085 0.84771659 0.774364   0.79009025]

F1 Mean:0.799490063085

Poly: Scores: [0.7012987, 0.63157895, 0.73611111, 0.73611111, 0.75   ]

Mean: 0.711019974178

F1 Scores: [0.70061622, 0.6389158, 0.74710089, 0.732282, 0.74301994]

F1 Mean: 0.712386970787

RBF: Scores: [0.41558442 0.48684211 0.40277778 0.375   0.375   ]

Mean: 0.411040859725

F1 Scores: [0.40318841 0.46642951 0.42463092 0.38231183 0.36566169]

F1 Mean: 0.408444472578

C. **Random Forest 5 Fold Results:**
Scores:[0.81818182 0.64473684 0.73611111 0.73611111 0.75   ]

Mean:0.737028176502

F1 Scores:[0.74594982 0.77619359 0.79129191 0.73495214 0.80150312]

F1 Mean:0.769978117359

2. Results without reduction, using SVM, RF, DT classifiers.

Initial shape of dataset: (369, 199325)

A. Decision Tree 5 Fold Results:

Scores:[0.64935065 0.65789474 0.66666667 0.68055556 0.72222222]

Mean:0.675337966127

F1 Scores:[0.68293423 0.67138004 0.70656977 0.73974451 0.70547788]

F1 Mean:0.701221287246

B. SVM 5 Fold Results for kernel :

linear

Scores:[0.61038961 0.57894737 0.625   0.61111111 0.65277778]

Mean:0.61564517354

F1 Scores:[0.62019704 0.5780541   0.63516484 0.61221376 0.6576092]

F1 Mean:0.620647785786

poly

Scores:[0.61038961 0.55263158 0.66666667 0.61111111 0.65277778]

Mean:0.618715348979

F1 Scores:[0.60658785 0.54944241 0.6700062   0.62020942 0.65634281]

F1 Mean:0.620517737211

rbf

Scores:[0.20779221 0.22368421 0.23611111 0.20833333 0.25   ]

Mean:0.225184172553

F1 Scores:[0.1017316   0.11805627 0.12179487 0.11293341 0.15277931]

F1 Mean:0.121459090555

C. Random Forest 5 Fold Results:

Scores:[0.62337662 0.59210526 0.72222222 0.625   0.61111111]

Mean:0.634763043974

F1 Scores:[0.68338429 0.55754579 0.68796512 0.6925062   0.55393413]

F1 Mean:0.635067106769

3. Using 80% of data for training and 20% as unknown for testing

Number of rows: 295

Size of data after feature reduction using decision tree: (295, 27)

A. Decision Tree 5 Fold Results:

```
Scores:[ 0.86666667  0.79661017  0.83050847  0.84745763  0.82758621]
```

Mean:0.83376582895

B. SVM 5 Fold Results: (linear kernel)

```
Scores:[ 0.7         0.71186441  0.76271186  0.76271186  0.77586207]
```

Mean:0.742630040912

C. Random Forest 5 Fold Results:

```
Scores:[ 0.81666667  0.72881356  0.76271186  0.81355932  0.84482759]
```

Mean:0.793315799727

Predicting over the remaining 20% data

A. Accuracy using Decision Tree on the Test Data: 0.648648648649

Confusion Matrix for Decision Tree on the Test Data:

```
[[11  3  0  2  0]
 [ 1  8  1  1  1]
 [ 1  4 11  2  0]
 [ 0  3  3  4  2]
 [ 1  1  0  0 14]]
```

B. Accuracy using SVM on the Test Data: 0.77027027027

Confusion Matrix for SVM on the Test Data:

```
[[15  1  0  0  0]
 [ 0 11  0  1  0]
 [ 0  1 13  4  0]
 [ 0  2  2  8  0]
 [ 2  2  1  1 10]]
```

C. Accuracy using Random Forest on the Test Data: 0.783783783784

Confusion Matrix:

```
[[15  1  0  0  0]
 [ 2 10  0  0  0]
 [ 0  2 14  2  0]
 [ 1  1  3  6  1]
 [ 0  2  1  0 13]]
```

Additional Observations

1. Characteristics of the selected 34 features :

Name	Average Length	Average EffectiveLength	Average TPM	Average NumReads
ENST00000436226.1	571	441.8793089	0.495906041	5.032292316
ENST00000636815.1	1724	1503.247046	0.032988772	1.1712757
ENST00000493165.1	813	387.1096369	2.577941726	29.37474268
ENST00000393657.6	1717	1712.740244	12.8298358	505.5403875
ENST00000260442.3	1249	873.7070894	10.29701831	202.4738905
ENST00000261210.9	820	680.6459837	1.779280445	27.96769163
ENST00000521270.5	581	377.6361409	0.149417094	1.443695176
ENST00000354454.7	2504	2553.794444	25.50681938	1492.718249
ENST00000480504.1	783	249.6599133	0.002088127	0.039864564
ENST00000451405.1	817	672.4773442	0.036498959	0.546425072
ENST00000591956.1	556	379.1562412	0.956490989	8.264067263
ENST00000567352.1	507	293.7908699	0.46909403	3.253335246
ENST00000456182.5	2411	2301.484119	4.763895965	242.960378
ENST00000480798.1	930	717.3463089	0.990480465	17.05420054
ENST00000615497.4	2383	2436.535528	11.16477829	626.2362141

ENST00000550772.1	429	329.8585772	2.437102531	17.85270637
ENST00000567491.1	1604	1149.822978	1.045427756	28.54569344
ENST00000382788.7	5512	5613.993279	0.496347954	63.29691341
ENST00000273063.10	4246	1261.888076	0.001559004	0.011457881
ENST00000500112.1	2131	1857.576856	0.223442638	9.403493388
ENST00000490410.1	3283	1860.433062	0.039779425	1.323301469
ENST00000583753.5	3560	3349.054878	0.147743561	11.57382822
ENST00000624581.1	1461	1341.340325	1.378479358	42.38278561
ENST00000339464.8	3919	3218.134661	0.166088627	12.96908442
ENST00000440428.5	1164	925.4655745	5.475653431	119.0525309
ENST00000558276.7	727	421.461607	1.829302552	17.99787883
ENST00000588073.1	760	667.9401328	0.677224397	10.24774865
ENST00000295326.4	444	256.6382818	21.14384775	124.1736076
ENST00000430640.1	414	208.5623713	0.56086137	3.18699187
ENST00000317269.7	3330	3212.614363	18.70806721	1390.46852
ENST00000425966.6	1684	1462.033713	0.405168826	13.64205859
ENST00000509893.2	1163	771.5876775	0.283209772	5.300813008
ENST00000527869.6	481	276.7609702	0.449203729	3.069454015
ENST00000306320.9	3268	2637.641572	1.318398746	77.70621616

We found that running the decision tree classifier multiple times in order to select the features, we get different set of features (But the count is same). The initial features (root and its immediate children) are same in all the results, but the nodes in the lower part of the tree keep changing. We think that this occurs due to multiple correlated features existing in the data set - The decision tree selects one of these features randomly.

Code submitted:

1. **5FoldCrossVal_ReducedFeatures.py:**
   Uses Decision tree to reduce features and applies Random Forest, SVM and Decision Tree classifiers.

2. **5FoldCrossValAllFeatures.py**
   Applies Random Forest, SVM and Decision Tree classifiers to the original features.

3. **20percent_Unknown.py**
   Trains the classifiers on 80% data after feature reduction and tests them on the 20% unknown data

4. **ProcessData.py**
   Converts input dataset to csv format which can be processed by sklearn.

5. **DetailsOfSelectedFeatures.py**
   Gets the average of Length, Effective Length, TPM and NumReads for the selected features.