IITM-CS5691 : Pattern Recognition and Machine Learning | Release Date: 23 Oct, 2021
New Method Writeup | Due Date: 10 Nov, 2021

Roll No: CS21M003_IC35020 | Name: Akash Saini & Returaj Burnwal

- Dear Student, You may have tried different methods for predicting each of the clinical descriptors in the data contest. Submit a write-up of the methods chosen for the data contest in the template provided below. **You will have to add the details in your own words and submit it as a team in gradescope**.
- We will run plagiarism checks on codes/write-up, and any detected plagiarism in writing/code will be strictly penalized.

1. (points) [Name of the Method chosen for each descriptor, and its Paradigm: paradigm could be linear/non-linear models, etc.)]:

Solution:

Descriptor : CO1
Model: SVM with rbf-kernel on 40 gene features.
Paradigm: Non-Linear Model
SkleanCode: SVC(kernel="rbf", C=0.5, class_weight="balanced")
Description: In the given dataset there were 22,282 features, we first selected 40 best gene features using analysis of variance F test. We also found that CO1 dataset was imbalanced so we used SVM-rbf kernel with class-weights equal to balanced.
In SVM, setting class-weight parameter equal to balanced is equivalent to using more regularization(parameter C in SVM) on less frequent class label. Using more regularization on some class label, forces the model to predict this class label more accurately.

Descriptor : CO2
Model: Gaussian Naive Bayes with 20 gene features
Paradigm: Non-Linear Model
SkleanCode: GaussianNB()
Description: In the given dataset there were 22,282 features, we first selected 20 best gene features using Analysis of Variance F test. We also found that CO2 dataset was slightly imbalanced. We over-sampled the less frequent class label dataset to address class imbalance.

Descriptor : CO3
Model: SVM with rbf with 35 gene features
Paradigm: Non-Linear Model

SkleanCode: SVC(kernel="rbf", C=0.9, class_weight={1:2.0, 0:0.7})
Description: In the given dataset there were 54,674 features, we first selected 35 best gene features using Analysis of Variance F test. We also found that CO3 dataset was highly imbalanced, so we had set higher regularization weight(parameter C in SVM) for less frequent class.

Descriptor : CO4
Model: Gaussian Naive Bayes with gene features whose variance is greater than 3.5
Paradigm: Non-Linear Model
SkleanCode: GaussianNB()
Description: In the given dataset there were 54,674 features, we first selected gene features whose variance is greater than 3.5. We also found that CO3 dataset was highly imbalanced, so we over-sampled the less frequent class label dataset to address class imbalance.

Descriptor : CO5
Model: SVM with rbf with 10 gene features
Paradigm: Non-Linear Model
SkleanCode: SVC(kernel="rbf", C=1.0, class_weight="balanced")
Description: In the given dataset there were 54,674 features, we first selected 10 best gene features using Analysis of Variance F test. We also found that CO5 dataset was slightly imbalanced, so we used SVM-rbf kernel with class-weight parameter equal to balanced.

Descriptor : CO6
Model: Linear discriminant with 10 gene features
Paradigm: Linear Model
SkleanCode: LinearDiscriminantAnalysis()
Description: In the given dataset there were 54,674 features, we first selected 10 best gene features using Analysis of Variance F test. We also found that CO6 dataset was imbalanced, so we over-sampled the less frequent class label dataset to address class imbalance.

2. (points) [Brief description on the dataset: could show graphs illustrating data distribution or brief any additional analysis/data augmentation/data exploration performed]

Solution:

Descriptor : CO1
a) CO1 dataset is imbalanced with number of samples for label_0 = 97 and label_1 = 33
b) We found that top 30 gene features had more than 30.0 analysis of variance (ANOVA) f test

score.
c) We tried to use PCA to reduce the dimension and we found that all the features in the reduced dimension had high singular values. We were not able to find a good model using all the PCA features.
d) We tried exploring with different SVM kernels and we found that SVM with rbf kernel with class_weights parameter equal to balanced gave better matthews correlation coefficient score.

Descriptor : CO2
a) CO2 dataset is slightly imbalanced with number of samples for label_0 = 77 and label_1 = 53
b) We found that top 25 gene features had more than 20.0 ANOVA-f test score.
c) Since the dataset was imbalanced, we over-sampled the less frequent label dataset.

Descriptor : CO3
a) CO3 dataset is imbalanced with number of samples for label_0 = 257 and label_1 = 83
b) We found that top 25 gene features had more than 30.0 ANOVA-f test score.
c) We explored different SVM kernels and we found that SVM with rbf with higher regularization weight(parameter C in SVM) for less frequent class gave better matthews correlation coefficient score.

Descriptor : CO4
a) CO4 dataset is imbalanced with number of samples for label_0 = 289 and label_1 = 51
b) To address class imbalance problem we oversampled the less frequent class label.

Descriptor : CO5
a) CO5 dataset is slightly imbalanced with number of samples for label_0 = 194 and label_1 = 146
b) We found that top 10 gene features had more than 300.0 ANOVA-f test score.

Descriptor : CO6
a) CO6 dataset is imbalanced with number of samples for label_0 = 140 and label_1 = 200
b) As most of the gene features had almost similar ANOVA-f test score values, we were not able to find significant gene features.

3. (points) [Brief introduction/motivation: Describe briefly the reason behind choosing a specific method for predicting each of the descriptors (could show plots or tables summarizing the scores of training different model)]

Solution:

a) We used Stratified 5 Fold cross validation to find average matthews correlation coefficient (MCC) score of the model. We selected the model which gave higher MCC score.

b) In Sklearn Parameters column, we have used default values for all the other parameters.

c) Top k Univariate features means top k gene features are selected using analysis of variance F test (ANOVA-F) score. A good feature will have higher ANOVA F-test score.

d) For most of the model we manually tuned the hyper-parameters.

e) Here we are reporting some of the models that have performed well relative to the other models that we have explored. We have added the complete list of explored models in our code.

Descriptor : CO1

Algorithm	**Handling Data Imbalance**	**Dimensionality Reduction**	**Sklearn Parameters**	**Average MCC Score**
Logistic Regression	Oversampling less frequent class label	Top 5 univariate gene features	solver="liblinear", penalty="l1"	0.391
Naive Bayes	Oversampling less frequent class label	Top 45 univariate gene features		0.496
Linear Discriminant Analysis	Oversampling less frequent class label	Gene features whose variance is greater than 0.5		0.471
Random Forest	Oversampling less frequent class label	Top 60 univariate features	criterion="entropy", max_depth=2, max_features=0.5	0.449
Gradient Boosting	Oversampling less frequent class label	Top 65 univariate features	max_depth=2, max_features=0.5, learning_rate=0.7	0.483
SVM	set high regularization value on less frequent class	Top 40 univariate features	kernel="rbf", class_weight ="balanced"	0.532

We choose, SVM with rbf kernel and 40 univariate features because it gave highest MCC score (0.532) among all the other models we have explored.

Descriptor : CO2

Algorithm	**Handling Data Imbalance**	**Dimensionality Reduction**	**Sklearn Parameters**	**Average MCC Score**
Logistic Regression	Oversampling less frequent class label	Projected the data to 10 principal component space	solver="liblinear", penalty="l1"	0.427
Naive Bayes	Oversampling less frequent class label	Top 20 univariate gene features		0.508
Linear Discriminant Analysis	Oversampling less frequent class label	Top 10 univariate gene features		0.445
Random Forest	Oversampling less frequent class label	Top 50 univariate features	criterion="entropy", max_depth=2, max_features=0.5	0.498
Gradient Boosting	Oversampling less frequent class label	Top 15 univariate features	max_depth=2, max_features=0.5, learning_rate=0.5	0.472
SVM	set high regularization value on less frequent class	Top 15 univariate features	kernel="poly", degree=3, class_weight ="balanced"	0.551

We choose, SVM with 3rd degree polynomial kernel and 15 univariate features because it gave highest MCC score (0.551) among all the other models we have explored.

Descriptor : CO3

Algorithm	**Handling Data Imbalance**	**Dimensionality Reduction**	**Sklearn Parameters**	**Average MCC Score**
Logistic Regression	Oversampling less frequent class label	Projected the data to 40 principal component space	solver="liblinear", penalty="l1"	0.316
Naive Bayes	Oversampling less frequent class label	Gene features whose variance is greater than 2.0		0.294
Linear Discriminant Analysis	Oversampling less frequent class label	Gene features whose variance is greater than 2.0		0.329
Random Forest	Oversampling less frequent class label	Top 30 univariate features	criterion="entropy", max_depth=2, max_features=0.5	0.310
AdaBoost	No processing performed	Top 25 univariate features	learning_rate=0.1	0.324
SVM	set high regularization value on less frequent class	Top 35 univariate features	kernel="rbf", C=0.9, class_weight ={1:2.0, 0:0.7}	0.354

We choose, SVM with rbf kernel and 35 univariate features because it gave highest MCC score (0.354) among all the other models we have explored.

Descriptor : CO4

Algorithm	**Handling Data Imbalance**	**Dimensionality Reduction**	**Sklearn Parameters**	**Average MCC Score**
Logistic Regression	Oversampling less frequent class label	Gene features whose variance is greater than 3.5	solver="liblinear", penalty="l1"	0.130
Naive Bayes	Oversampling less frequent class label	Gene features whose variance is greater than 3.5		0.218
Linear Discriminant Analysis	Oversampling less frequent class label	Gene features whose variance is greater than 2.5		0.205
AdaBoost	No processing performed	Top 25 univariate features	n_estimators=200, learning_rate=0.1	0.231

We choose, Naive Bayes and gene features whose variance is greater than 3.5, because it has more stable MCC score (0.218).

Descriptor : CO5

Algorithm	**Handling Data Imbalance**	**Dimensionality Reduction**	**Sklearn Parameters**	**Average MCC Score**
Logistic Regression	Oversampling less frequent class label	Top 10 univariate features	solver="liblinear", penalty="l1"	0.84
SVM	set high regularization value on less frequent class	Top 10 univariate features	kernel="rbf", C=1.0, class_weight ="balanced"	0.84

Here both algorithms performed equally well. We here choose SVM with rbf kernel and 10 univariate features.

Descriptor : CO6

Algorithm	**Handling Data Imbalance**	**Dimensionality Reduction**	**Sklearn Parameters**	**Average MCC Score**
Linear Discriminant Analysis	Oversampling less frequent class label	Top 10 univariate features		0.122
Random Forest	Oversampling less frequent class label	Top 30 univariate features	criterion="entropy" max_depth=2, max_features=0.5	0.113
Gradient Boosting	Oversampling less frequent class label	Top 25 univariate features	max_depth=1, subsample=0.5 max_features=0.5, learning_rate=0.5	0.142

We choose Linear Discriminant Analysis and 10 univariate features because it has more stable MCC score (0.122).

4. (points) [As the data comes from a clinical setting, model interpretation is also an important task along with prediction. Describe briefly your preferred choice of methods if you are asked to determine the significant genes behind the regulation of the endpoints or restrict the number of features/genes.]:

Solution:

We tried to reduce the dimension of feature space using:
a) PCA : Principal Component Analysis
b) Variance Threshold: Select feature only when its variance is greater than some threshold. c) analysis of variance F test (ANOVA-F) score: Good features have high score

We found that using ANOVA-F test score we were able to get better classifiers in most of our descriptors. We would use ANOVA-F test scores to determine significant genes.

We have also reported the number of significant genes for each descriptor in the answer of question 3.

5. (points) [Share your thoughts on each of the endpoints: whether easy/difficult to predict]:

Solution:

Descriptor : CO1
We were able to achieve matthews correlation coefficient (MCC) score of 0.532 on this class using SVM. It was of medium difficulty.

Descriptor : CO2
We were able to achieve MCC score of 0.508 using Naive Bayes. It was of medium difficulty. Algorithms like adaboost, random forest, gradient boosting sometimes gave better MCC score but their MCC scores had high fluctuations if we change the training and testing dataset.

Descriptor : CO3
We were able to achieve MCC score of 0.354 using SVM. It was difficult to predict. We were not able to improve the MCC score above 0.354. Calibrating SVM paramters were challenging. Similar to CO2, algorithms like adaboost, random forest, gradient boosting sometimes gave better MCC score but their MCC scores had high fluctuations if we change the training and testing dataset.

Descriptor : CO4
We were able to achieve MCC score of 0.218 using Naive Bayes. It was difficult to predict.

Descriptor : CO5
We were able to achieve MCC score of 0.837 using SVM. It was easy to predict. We were able to find significant gene features and training only on these features gave MCC score of 0.837.

Descriptor : CO6
We were able to achieve MCC score of 0.122 using linear discriminant analysis. It was very difficult to predict. We explored many different models like SVM, random forest, gradient boosting, logistic regression, adaboost but none of the model gave stable MCC score above 0.122.

6. (points) [Add any additional information: like challenges faced or some details that would help us to better understand the strategies that you have utilized for model development]:

Solution:

Challenges Faced:
a) Both training dataset_1 and dataset_2 had very large number of genes' information and very small number of samples. This made the classification task difficult because most of our ML

algorithms easily overfit the training dataset. Finding good significant gene features was challenging.

b) Most of the descriptor's dataset was imbalanced. We explored different ways to address imbalance problem either by over-sampling/under-sampling or by regularizing more less frequent class label.

c) Finding a good classifier for descriptor CO6 was really challenging.

d) Finding a good classifier for descriptor CO3 and CO4 was also difficult.