Kidney Tissue Classification with One vs All Method Using Gene Expression Levels of Microarray Data

Furkan Kılıçaslan

Abstract

In this experiment, several feature selection and classification methods were used to assess tissue type from gene expression levels of 50 different genes.

Keywords: microarray, classification, kidney, feature selection, gene expression

1. **Introduction**

Machine learning has played a crucial role in aiding doctors with diagnosis procedures. Various classification models are in use in this role to improve the decisions of doctors in diagnosis.

2. **Experiment Goal**

This experiment approaches gene expression data from a microarray experiment to build a classification model that can classify kidney tissue against other tissues. Also it aims to find the most relevant genes for this type of classification by using different feature selection methods.

3. **Data Preparation**

Data is loaded into a jupyter notebook working space using pandas [1] library. Then the target label values are converted to binary. The obtained data is scaled to have unit mean and variance using scikit-learns StandardScaler() [2]. After this step the training data we have which has 100 samples have been split into two using scikit-learn train_test_split() [2] with a 70/30 split.

4. Feature Selection

For the feature selection process, Recursive Feature Elimination has been used to obtain the most relevant 20 features from the total 50 features. The index locations of these features as follows :

[1, 2, 4, 5, 7, 8, 9, 11, 19, 20, 23, 29, 30, 31, 33, 34, 36, 38, 43, 45]

5. Model Selection

Here in this part three different classification models were picked and applied on the dataset. Decision Trees, Random Forest Classifiers, Logistic Regression. The obtained classification reports and confusion matrices are as follows in order :

Decision Trees

```
precision   recall  f1-score   support                    [[24  1]
                                                           [ 1  4]]
       0      0.96      0.96      0.96       25
       1      0.80      0.80      0.80        5

avg / total      0.93      0.93      0.93       30
```

Random Forest Classifier

```
precision   recall  f1-score   support                    [[25  0]
                                                           [ 1  4]]
       0      0.96      1.00      0.98       25
       1      1.00      0.80      0.89        5

avg / total      0.97      0.97      0.97       30
```

Logistic Regression

```
precision   recall  f1-score   support                    [[23  2]
                                                            [ 0  5]]
       0      1.00      0.92      0.96       25
       1      0.71      1.00      0.83        5

avg / total      0.95      0.93      0.94       30
```

6. Conclusion

For this experiment and the data it is clear that the Random Forest Classifier produces the most robust results and the highest classification rate for kidney tissue which is our most important target label. Even though the sample amount is low, having a high dimensionality of 50 different gene expression levels proved useful to obtain a high classification score in classifying kidney tissue against other tissue.

The results achieved from this experiment can further be improved by applying more rigorous feature selection and data preprocessing methods as well as obtaining more training data. The predicted results for our test data is as below : (1 for kidney 0 for other tissue)

```
array([0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0,0, 0, 0, 0, 0, 0, 1, 0])
```

References

1) McKinney, Wes (2018, May 1) pandas library. Retrieved from https://pandas.pydata.org/about.html

2) Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.