Title: Classifying cancerous tissue using gene expression data

Introduction

DNA microarray technology is used to measure the expression levels of genes under various conditions. Differential expression of these genes can elucidate newer insights into the functioning of various diseases. Usually, the problem faced in microarray expression analysis is that of high dimensionality. There are a lot more genes than samples available for analysis. So, in this project we decided to identify several datasets that suffer from the curse of dimensionality and apply various feature engineering techniques to reduce the number of genes in each case. Finally, we will build classifiers on the top of each of these feature-reduced datasets and compare their accuracies.

Literature search for multiple microarray data

The following four datasets were considered for our analysis

Name	Disease type	Number of features	Number of samples	Number of classes
Alon (1999)	Colon	1977	62	2
Gravier (2010)	Breast cancer	2905	168	2
Gordon (2002)	Lung cancer	12533	181	2
Golub (1999)	Leukemia	7129	72	2

Dataset 1: *Gravier et al*

Gravier et al. (2010) have considered small, invasive ductal carcinomas without axillary lymph node involvement (T1T2N0) to predict metastasis of small node-negative breast carcinoma. Using comparative genomic hybridization arrays, they examined 168 patients over a five-year period. The 111 patients with no event after diagnosis were labelled good, and the 57 patients with early metastasis were labelled poor.

Analysis

Two different feature selection approaches were tried:

1. L1 regularized logistic regression

2. Recursive feature elimination

The following workflow was undertaken:

1. Split the dataset into train and test
2. Apply the two feature selection methods on the train set only and select the top features. Let's call this the feature reduced train set. Similarly, select the top features from the test set. Let's call this the feature reduced test set.
3. Since this is an imbalanced dataset, use SMOTE to synthetically sample minority class data. SMOTE is applied on the train data only
4. Now for each of the different classifiers (SVM, Random forests, Decision trees), do parameter tuning using the train data only
5. Now fit the model using the tuned parameters on the feature reduced train set.
6. Apply the fitted model on the feature reduced test set and report the classification metrics.

Results (Recursive feature elimination)

SVM

Class	**Precision**	**recall**	**F1 score**	**AUROC**
No metastasis	0.75	0.88	0.81	0.647
Metastasis	0.64	0.41	0.50	

Random forest

Class	**Precision**	**recall**	**F1 score**	**AUROC**
No metastasis	0.80	0.82	0.81	0.705
Metastasis	0.62	0.59	0.61	

Decision trees

Class	**Precision**	**recall**	**F1 score**	**AUROC**
No metastasis	0.76	0.74	0.75	0.705
Metastasis	0.50	0.53	0.51	

Results (LASSO based feature selection: 63 out of 2842 features chosen)

SVM

Class	**Precision**	**recall**	**F1 score**	**AUROC**
No metastasis	0.79	0.76	0.78	0.676
Metastasis	0.56	0.59	0.57	

Random forest

Class	Precision	recall	F1 score	AUROC
No metastasis	0.86	0.88	0.87	0.797
Metastasis	0.75	0.71	0.73	

Decision trees

Class	Precision	recall	F1 score	AUROC
No metastasis	0.78	0.85	0.82	0.794
Metastasis	0.64	0.53	0.58	

Dataset 2: Gordon et al

Using microarray data for the pathological distinction between malignant pleural mesothelioma (MPM)and adenocarcinoma (ADCA) of the lung. There were 181 tissue samples (31 MPM and 150 ADCA).

Analysis

Two different feature selection approaches were tried:

3. L1 regularized logistic regression
4. Recursive feature elimination

The following workflow was undertaken:

7. Split the dataset into train and test
8. Apply the two feature selection methods on the train set only and select the top features. Let's call this the feature reduced train set. Similarly, select the top features from the test set. Let’s call this the feature reduced test set.
9. Since this is an imbalanced dataset, use SMOTE to synthetically sample minority class data. SMOTE is applied on the train data only
10. Now for each of the different classifiers (SVM, Random forests, Decision trees), do parameter tuning using the train data only
11. Now fit the model using the tuned parameters on the feature reduced train set.
12. Apply the fitted model on the feature reduced test set and report the classification metrics.

Results (LASSO based feature selection - 35 selected out of 12533)

SVM

Class	Precision	recall	F1 score	AUROC
ADCA	0.98	1.00	0.99	0.944
MPM	1.00	0.89	0.94	

Random forest

Class	Precision	recall	F1 score	AUROC
ADCA	0.98	0.98	0.98	0.9335
MPM	0.89	0.89	0.89	

Decision trees

Class	Precision	recall	F1 score	AUROC
ADCA	0.98	1.00	0.99	0.9335
MPM	1.00	0.89	0.94	

Results (Recursive feature elimination)

SVM

Class	Precision	recall	F1 score	AUROC
ADCA	0.98	0.98	0.98	0.9335
MPM	0.89	0.89	0.89	

Random forest

Class	Precision	recall	F1 score	AUROC
ADCA	0.98	1.00	0.99	0.944
MPM	1.00	0.89	0.94	

Decision trees

Class	Precision	recall	F1 score	AUROC
ADCA	0.94	0.96	0.95	0.9444
MPM	0.75	0.67	0.71	

Dataset 3: Golub et al

The Golub et al. (1999) data consist of 47 patients with acute lymphoblastic leukemia (ALL) and 25 patients with acute myeloid leukemia (AML). Each of the 72 patients had bone marrow samples obtained at the time of diagnosis. Furthermore, the observations have been assayed with Affymetrix Hgu6800 chips, resulting in 7129 gene expressions (Affymetrix probes).

Results (LASSO based feature selection - 25 selected out of 7129)

SVM

Class	Precision	recall	F1 score	AUROC
ALL	0.82	1.00	0.90	0.8125
AML	1.00	0.62	0.77	

Random forest

Class	Precision	recall	F1 score	AUROC
ALL	0.93	1.00	0.97	0.9375
AML	1.00	0.88	0.93	

Decision trees

Class	Precision	recall	F1 score	AUROC
ALL	0.93	0.93	0.93	0.9375
AML	0.88	0.88	0.88	

Results (Recursive feature elimination)

SVM

Class	Precision	recall	F1 score	AUROC
ALL	0.93	1.00	0.97	0.9375
AML	1.00	0.88	0.93	

Random forest

Class	Precision	recall	F1 score	AUROC
ALL	0.93	1.00	0.97	0.9375
AML	1.00	0.88	0.93	

Decision trees

Class	Precision	recall	F1 score	AUROC
ALL	0.87	0.93	0.90	0.9375
AML	0.86	0.75	0.80	

Dataset 4: Alon et al

Alon et al. (1999) have presented a data set that contains gene expression levels of 40 tumour and 22 normal colon tissues for 6500 human genes obtained with an Affymetrix oligonucleotide array.

Results (LASSO based feature selection - 23 selected out of 1977)

SVM

Class	Precision	recall	F1 score	AUROC
Normal	0.83	0.71	0.77	0.815
Cancerous	0.85	0.92	0.88	

Random forest

Class	Precision	recall	F1 score	AUROC
Normal	0.78	1.00	0.88	0.9166
Cancerous	1.00	0.83	0.91	

Decision trees

Class	Precision	recall	F1 score	AUROC
Normal	0.88	1.0	0.93	0.9166
Cancerous	1.00	0.92	0.96	

Results (Recursive feature elimination based feature selection)

SVM

Class	Precision	recall	F1 score	AUROC
Normal	0.60	0.43	0.50	0.630
Cancerous	0.71	0.83	0.77	

Random forest

Class	Precision	recall	F1 score	AUROC
Normal	0.86	0.86	0.86	0.886
Cancerous	0.92	0.92	0.92	

Decision trees

Class	Precision	recall	F1 score	AUROC
Normal	0.50	1.0	0.67	0.886
Cancerous	1.00	0.42	0.59	

Discussion

In this brief analysis we analyzed four different datasets with high dimensions and applied two different feature selection methods to reduce dimensionality. We achieved reasonable classification accuracy in terms of ROC score, precision, recall and f1-score. Overall, LASSO followed by tree based methods consistently ranked better as compared to others.
Further analysis with more datasets need to be done. Also, the results reported here must be compared with other existing benchmark studies.