Evaluating Performance of Regression and Classification Models Using Known Lung Carcinomas Prognostic Markers

Shrikant Pawar[1(✉)], Karuna Mittal[2], and Chandrajit Lahiri[3]

[1] Department of Computer Science and Biology, Claflin University, Orangeburg, USA
spawar@claflin.edu
[2] Department of Biomedical Sciences, Emory University, Atlanta, USA
[3] Department of Biological Sciences, Sunway University, Petaling Jaya, Malaysia
chandrajitl@sunway.edu.my

Abstract. Differential expression study between tumor and non-tumor cells aids lung cancer diagnostic classifications and prognostic prediction at various stages. Support vector machine (SVM) learning is used to categorize the morphology of lung cancer. Logistic regression, random forest, and group lasso-based models are used to model dichotomous outcome variables. The purpose is to take groups of observations and design boundaries to forecast which group future observations belong to base measurements. The performance of these selected regression and classification models using lung cancer prognostic indicators is evaluated in this article. The presented results might guide for further regularizations in classification techniques using known lung carcinoma marker genes.

Keywords: Regression · Lung carcinomas · Predictions

1 Introduction

Among all malignancies, lung cancer caused the most considerable loss of pay, totaling $21.3 billion in year 2018–19 [1]. However, the specific environmental and genetic etiology of a person's lung cancer is unknown, and it can be described as a tumor forming in the lung when altered cells escape the immune system and grow out of control. Despite the fact that many lung cancer research findings have been published, scientific advancement in lung cancer research is still limited. Lung cancer diagnostic classifications and prognosis prediction at various stages are aided by differential expression analysis between tumor and non-tumor cells. Attempts have been undertaken to find genes linked to lung cancer symptoms. Lung cancer morphology categorization has been performed using support vector machine learning techniques [2]. Alanni et al. devised a deep gene selection technique for cancer classification from microarray datasets [3]. The results of their experiments revealed an average sensitivity of 95.22% and a specificity of 77.39%. Several machine learning methods have also been utilized to identify 13 top genes in lung adenocarcinoma and lung squamous cell cancer [4]. To learn cancer

© Springer Nature Switzerland AG 2022
I. Rojas et al. (Eds.): IWBBIO 2022, LNBI 13347, pp. 413–418, 2022.
https://doi.org/10.1007/978-3-031-07802-6_35

type classification based on TCGA data, Mohammed et al. employed the least absolute shrinkage and selection operator (LASSO) as a feature selection approach [5]. In addition to cancer classification and biomarker identification, overlapping feature selection strategies have been used [6]. Squamous cell lung cancer (LUSC) has been associated to four genes CCNA2 (890), AURKA (6790), AURKB (9212), and FEN1 (2237) [7], while lung adenocarcinoma (LUAD) has been linked to four genes (CD44 (960), CCND3 (896), NCALD (83988), MACF1 (23499), and RAMP2-AS1 (10266). In a comprehensive genomic study of squamous cell lung tumors [9], one gene, TP53 (7157), was found to be altered in virtually all cases. To model dichotomous outcome variables, logistic regression, random forest, support vector machines (SVM), and group lasso-based models are utilized [10, 11]. The purpose is to take groups of observations and design boundaries to forecast which group future observations belong to base on their measurements. The performance of these selected regression and classification models using lung cancer prognostic indicators is evaluated in this article.

2 Dataset and Methodology

We chose to test performance of each of the 4 techniques on 3 different datasets with lung LUAD (517 tumor, 59 normal) [12], LUSC (501 tumor, 51 normal) [9] and non-small cell lung carcinomas (NSCLC) (91 tumor, 65 normal subjects) [13]. Libraries randomForest, caret was used for random forest application, library kernlab and e1071 for SVM, and glmnet for regression. Functions svm(kernel = "radial", cost = 10, gamma = 1), predict(), glm(), wald.test(), and glmnet() were utilized for performing k-fold cross-validation to find optimal lambda value that minimizes test mean squared error (MSE) [14–16]. Cross validations were performed with 70:30 training to testing splits. Response value was considered 0/living and 1/death status. Sum of squares total (SST), sum of squares error (SSE) and R-squared value on a response variable (y) were calculated as follows:

sst < - sum((y-mean(y))^2).

sse < - sum((y_predicted-y)^2).

rsq < - 1 - sse/sst.

All the code for accessing data and methodology can be found at authors GitHub account: https://github.com/spawar2/Regression-Lung-Carcinoma/tree/main.

3 Results

3.1 Prediction Performance of Random Forest

Test classification accuracy of 55% was obtained on selected 10 genes expression values with an 30–78 range for 95% CI. The P value was seen insignificant with sensitivity and specificity of 14 and 81% respectively. The 10 genes were not found to exclusively classify the survival response status. We also tested this classification approach on different combinations of these 10 marker genes, and results were consistent. Table 1 provides details of test and training metrics of random forest.

Table 1. Test and training metrics of random forest.

	Train: Type of random forest: regression Number of trees: 500 No. of variables tried at each split: 3 Mean of squared residuals: 0.2553676 % Var explained: −6.72	Test: Type of random forest: classification Number of trees: 500 No. of variables tried at each split: 3 OOB estimate of error rate: 40.62%
Accuracy	1	0.5556
95% CI	(0.944, 1)	(0.3076, 0.7847)
No information rate	0.6094	0.6111
P-value [Acc > NIR]	1.709e−14	0.7680
Kappa	1	−0.0435
Sensitivity	1	0.14286
Specificity	1	0.81818
Pos pred value	1	0.33333
Neg pred value	1	0.60000
Prevalence	0.3906	0.38889
Detection rate	0.3906	0.05556
Detection prevalence	0.3906	0.16667
Balanced accuracy	1	0.48052
'Positive' class	0	0

3.2 Prediction Performance of SVM

Testing SVM with 10 marker gene expression on a survival response variable predicted 85% subjects living/0 correctly (n = 20), and 24% subjects dead/1 correctly (n = 62) (Table 2). The test group was randomly selected with Fig. 1 showing dispersion of 2 groups for genes 890 and 6790. We found similar dispersion patterns for other genes and throughout all the 3 separate datasets. SVM poorly classifies survival response status with known marker genes.

Table 2. SVM classification of test data.

	0	1
0	17	3
1	15	47

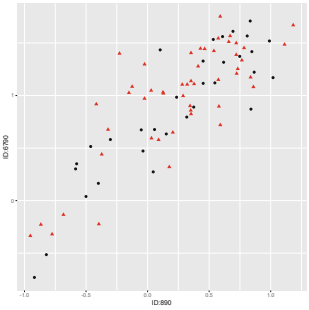

Fig. 1. Dispersion of survival and dead subjects for genes 890 and 6790.

3.3 Prediction Performance of Logistic Regression and LASSO

Testing prediction probabilities from LASSO ranged from 0.3–0.7 (Table 3). A weighted distance between the unrestricted estimate (Wald test) P value was found to be insignificant. The Chi-squared value of 0.89 with a P value > 0.05 also states insignificant prediction probabilities. The least squares regression tries to find coefficient estimates that minimize the sum of squared residuals (RSS). It can be presented with function: RSS $= \Sigma(y_i - \hat{y}_i)2$, y_i: is actual response value for the i^{th} observation and $\hat{\mathbf{y}}_\mathbf{i}$: is the predicted response value based on the multiple linear regression model. Figure 2 depicts calculates the binomial deviance (binomial log-likelihood) in the test dataset. The test data R square value of -6.70 was obtained stating the selected model does not follow the trend of the data, therefore leading to a worse fit than the horizontal line.

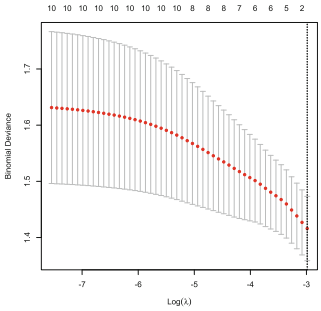

Fig. 2. Calculation of binomial deviance (binomial log-likelihood) in the test dataset.

Table 3. Prediction probabilities from LASSO.

Status	Predicted probability
0	0.6545150
0	0.6875263

(*continued*)

Table 3. (*continued*)

Status	Predicted probability
0	0.5171204
1	0.6935557
1	0.6536800
1	0.7114294
1	0.7818345
1	0.3633772
1	0.8720003
1	0.6644866
1	0.6111339
1	0.6527981

4 Discussion and Future Scope

The biological literature of the selected 10 key genes is enriched by their new roles associated to lung cancer, which have moved from an indirect to a direct association, i.e., to become new biomarkers. In many cases, indirect impacts are more important than direct effects because direct effects can be seen and controlled, whereas indirect effects are difficult to detect and control. We wanted to test their effects on response variable using selected regression and classification techniques. We find insignificant correlations with response variable. these findings are consistent for all the three cancer types. There can be several reasons of these outcomes. Growing more than one type of lung cancer is uncommon among all known lung cancer types. As a result, competing risk factor models can be extremely effective at modeling a variety of lung cancer forms. Further, confounding factors (age, gender, preexisting conditions, etc.) also significantly affect regression perditions. The expression data is rarely linearly separable, and prone to noise and overfitting. Although we did take care of limiting outliers, regression techniques are oversensitive to nominal outliers. One limitation of this study is multicollinearity, dimensionality reduction techniques are needed to be implemented to address issue of multicollinearity apart from above confounding factors. The presented results might guide for further regularizations in classification techniques using known lung carcinoma marker genes.

Author Contributions. SP and CL conceived the concepts, planned, and designed the article. SP and CL primarily wrote and edited the manuscript.

Funding Source. No external funding has been utilized for this study.

Competing Interests. The authors declare that they have no competing interests.

References

1. Islami, F., et al.: National and state estimates of lost earnings from cancer deaths in the united states. JAMA Oncol. **5**(9), e191460 (2019). https://doi.org/10.1001/jamaoncol.2019.1460
2. Podolsky, M.D., Barchuk, A.A., Kuznetcov, V.I., Gusarova, N.F., Gaidukov, V.S., Tarakanov, S.A.: Evaluation of machine learning algorithm utilization for lung cancer classification based on gene expression levels. Asian Pac. J. Cancer Prev. **17**(2), 835–838 (2016). https://doi.org/10.7314/apjcp.2016.17.2.835. PMID: 26925688
3. Alanni, R., Hou, J., Azzawi, H., Xiang, Y.: Deep gene selection method to select genes from microarray datasets for cancer classification. BMC Bioinformatics **20**(608), 1–15 (2019). https://doi.org/10.1186/s12859-019-3161-2
4. Yuan, F., Lu, L., Zou, Q.:. Analysis of gene expression profiles of lung cancer subtypes with machine learning algorithms. Biochim. Biophys. Acta (BBA)–Mol. Basis of Dis. **1866**(8), 165822 (2020). doi: https://doi.org/10.1016/j.bbadis.2020.165822. ISSN 0925–4439. https://www.sciencedirect.com/science/article/pii/S0925443920301678
5. Mohammed, M., Mwambi, H., Mboya, I.B., Elbashir, M.K., Omolo, B.: A stacking ensemble deep learning approach to cancer type classification based on TCGA data. Sci. Rep. **11**(1), 15626 (2021). https://doi.org/10.1038/s41598-021-95128-x
6. Chen, J.W., Dhahbi, J.: Lung adenocarcinoma and lung squamous cell carcinoma cancer classification, biomarker identification, and gene expression analysis using overlapping feature selection methods. Sci. Rep. **11**(1), 13323 (2021). https://doi.org/10.1038/s41598-021-92725-8
7. Gao, M., Kong, W., Huang, Z., Xie, Z.: Identification of key genes related to lung squamous cell carcinoma using bioinformatics analysis. Int. J. Mol. Sci. **21**(8), 2994 (2020). doi: https://doi.org/10.3390/ijms21082994. ISSN 1422-0067. https://www.mdpi.com/1422-0067/21/8/2994
8. Song, Z., Zhang, Y., Chen, Z., Zhang, B.: Identification of key genes in lung adenocarcinoma based on a competing endogenous RNA network. Oncol. Lett. **21**(1), 60 (2021). https://doi.org/10.3892/ol.2020.12322
9. Cancer Genome Atlas Research Network: Comprehensive genomic characterization of squamous cell lung cancers. Nature **489**(7417), 519–525 (2012). https://doi.org/10.1038/nature13385
10. Hosmer, D., Lemeshow, S.: Applied Logistic Regression, 2nd edn. John Wiley & Sons Inc., New York (2000)
11. Long, J.S.: Regression Models for Categorical and Limited Dependent Variables. Sage Publications, Thousand Oaks (1997)
12. Cancer Genome Atlas Research, Network: Comprehensive molecular profiling of lung adenocarcinoma. Nature **511**(7511), 543–550 (2014). https://doi.org/10.1038/nature13385
13. Hou, J., et al.: Gene expression-based classification of non-small cell lung carcinomas and survival prediction. PLoS ONE **5**(4), e10312 (2010). https://doi.org/10.1371/journal.pone.0010312
14. Breiman, L.: Random forests. Mach. Learn. **45**, 5–32 (2001). https://doi.org/10.1023/A:1010933404324
15. Karatzoglou, A.: kernlab–An S4 package for kernel methods in R. Kernel-Based Machine Learning Lab (2019)
16. Friedman, J.: Regularization paths for generalized linear models via coordinate descent. Lasso and Elastic-Net Regularized Generalized Linear Models (2009)