Classifying tumors with RNA-sequencing data using machine learning models

Anh Nguyen
u_504395
Group 18
Reasearch Workshop for CSAI
Department of Cognitive Science and Artificial Intelligence

Abstract

Over the last decade, RNA-sequencing data analysis has become a widely utilized method in the medical field. Specifically, it can be used to detect and distinguish different classes of tumor growth from gene expression. With RNA-sequencing data combined with machine learning (ML) algorithms, it is feasible to classify different types of tumors with high accuracy. This research aims to test the accuracy of different ML models on endothelial cell data.

Keywords: Blood-brain barrier; Glioblastoma; lung cancer brain metastasis; RNA-sequencing data; machine learning; Support Vector Machine; random forest, K-Nearest Neighbors.

Introduction

Cancer is a lethal disease that has existed since 3000 BC (American Cancer Society, 2018). Multiple researches are dedicated to the detection and treatment of cancerous tumors. In the past decade, through some significant advancements in the medical/technological field, various techniques have been implemented that allow us to gain additional insight into tumor growth.

Glioblastoma (GBM) is the most malignant and frequent primary brain tumor found in adults, while secondary tumors such as brain metastasis (BM) can be originated from extracranial parts such as breast, lung, or skin. Early detection and identification of these cancerous tumors are crucial for treatment and medication since it can drastically increase patients' survival rates. The blood-brain barrier (BBB) possesses unique properties that allow the regulation of nutrients and cells between the blood and the brain (Daneman, Prat, 2015). Alterations in BBB and endothelial cells in BBB would indicate the progression of a tumor (Tiwary et al., 2018). Several studies have previously incorporated ML algorithms to aid the classification of different categories of cancer. Specifically, classification models such as support vector machine (SVM) and Naive Bayes classifier were used to identify subclasses of breast cancer (Omondiagbe, Veeramani, Sidhu, 2019). Although prediction accuracy is heavily emphasized in the medical field, ML algorithms still have promising prospects for developing treatments for cancerous patients. Once these algorithms are refined, there will be instant identification of different classes of cancer, resulting in the improvement of patients' outcomes (Tan, Gilbert, 2003).

Method

Dataset

The endothelial cell dataset contains RNA-sequenced data, including labeled samples of 21908 gene expressions for 17 different subjects: 6 healthy controls (HC), 6 lung cancer brain metastasis (BM) and 5 glioblastoma (GBM). The transcriptional alteration of endothelial cells recorded will be used for classification against the criterion of the 17 classified examples of HC, BM and GBM.

Preprocessing steps

Log scaling was used to scale the dataset primarily because ML algorithms work better with Gaussian distributed data and prevent the curse of dimensionality from significantly altering the results. Another reason for using log transformation is that the features can hold drastically different value ranges. Thus, log scaling the dataset would aid in heavily reducing the bias of, potentially, a skewed dataset yet retain the significant differences between the values of the feature variables(endothelial cells).

Min-max scaling was another method used to preprocess the data and to be compared with log scaling. Similar to log scaling, it preserves the significant differences between the values of the features needed for ML classifiers to learn. However, Min-Max scaler has the limitation of not being weak for a data distribution with severe skewness and outliers.

Principal Component Analysis (PCA) was used to reduce the high dimensionality of the dataset, being the 21908 labeled and quantified gene expressions. By keeping 95 percent of the data variance, the most significant features were left to accommodate 95 percent of the variance found in the dataset.

Implementation

SVM, Random Forest and K-Nearest Neighbors (KNN) classifiers were tested and compared. SVM was selected since it is effective with high dimensional datasets and when the number of features exceeds the number of samples. For the alternative classifier, Random Forest was used mainly for the large scale of the dataset and its reliability than using single decision trees. KNN classifier was picked as a third test model to compare against the two previous models. Grid search was used to find the best parameter for each model by finding the highest mean leave-one-out cross-validation score.

Result

Figure 1 shows the performance of Linear Support Vector Classifier (SVC), K-Nearest Neighbor Classifier (KNN) and Random Forest Classifier on the log-scaled data and min-max scaled data as shown in the table. Overall, all models give better cross-validation scores with log-scaled data than with

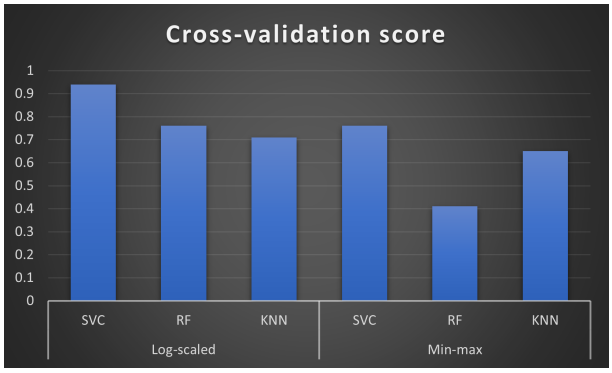

Figure 1: Cross-validation scores of Linear SVC, Random Forest and KNN classifier on log-scaled and min-max-scaled data

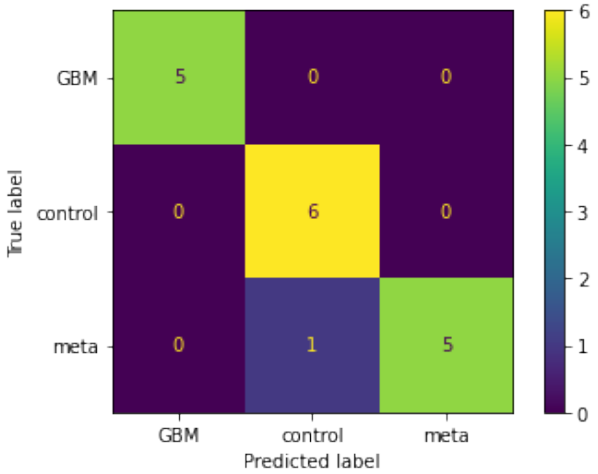

Figure 2: Cross-validation confusion matrix for Linear SVC (C = 0.01, max_iter = 10000) on log-scaled data

min-max-scaled data. Out of all the tested models, Linear SVC delivers the highest cross-validation score of 0.94 and 0.76 in both cases. KNN classifier and Random Forest classifier perform significantly worse than Linear SVC with 0.71 and 0.76 respective scores on log-scaled data. Both models score even lower on min-max scaled data, with 0.65 and 0.41 respectively for KNN and Random Forest classifier. Figure 2 illustrates the confusion matrix of Linear SVC's prediction on the log-scaled data. The best parameter for Linear SVC with log-scaled data is C = 0.01, max_iter = 10000. Out of 17 cross-validation tests, only one prediction error was found.

Discussion

Based on the result, it is apparent that standard ML algorithms can classify tumors using endothelial gene data with high accuracy. In general, the performance of every tested model on log-scaled data is higher than that of min-max-scaled data. It indicates that the decision to use log transformation on the dataset is reasonably justified and that machine learning models provide different results for different data processing steps. Linear SVC generating the highest cross-validation score of 0.94 shows that it can work well with this particular dataset. High classification accuracy is essential in the medical field since a prediction error could result in a patient's life. Therefore, it is not feasible to employ Random Forest and KNN classifier for this particular dataset since their cross-validation score is inadequate which might lead to errors when applied to other datasets.

There are, without a doubt, several limitations to this experiment. The most obvious one is the low sample size of this dataset. The sample size being 17 results in a less meaningful model accuracy since one participant's data can heavily impact the parameters. Although the dataset was preprocessed to allow for a low sample size, this issue needs to be accounted for when interpreting the outcome of this experiment. A larger dataset is required for better generalization of the model. Secondly, the model used for training in this study is a standard model from the sklearn library, meaning that these models might not be as sophisticated as other recently developed models such as neural networks. As a result, the cross-validation score is not ideal for all tested models.

Despite the drawbacks mentioned above, the findings of this experiment still align with the result of prior research, which is that machine learning algorithms are cheap and efficient in identifying cells with high accuracy. Researchers can compare and test more complex ML learning algorithms to classify tumors in the future. For example, a linear binary classifier that is capable of working with high-dimension, low-sample size data was recently designed (Shen, Er, Yin, 2020).

Conclusion

This paper analyzed the endothelial cells dataset using data scaling, dimensionality reduction methods and three prevalent ML algorithms to identify tumors.The experiment proves that ML models can predict gene labels efficiently with high accuracy. Results indicate that all three models perform better on log-scaled data than on min-max-scaled data. Additionally, Linear SVC has the highest cross-validation accuracy of 0.94 out of all algorithms. This research also shows that feature selection and hyperparameter tuning can assist in the diagnosis of malignant tumors when combined with ML techniques. Future works can aim towards improving the chosen method into a practical application for aiding doctors in identifying malignant cells. More complex ML algorithms can also be tested for a better cancer diagnosis.

References

American Cancer Society, . (2018). Understanding what cancer is: Ancient times to present. Retrieved from https://www.cancer.org

Daneman, R., & Prat, A. (2015). The blood-brain barrier. doi: https://doi.org/10.1101/cshperspect.a020412

Omondiagbe, D. A., Veeramani, S., & Sidhu, A. S. (2019).

Machine learning classification techniques for breast cancer diagnosis. *Material, Science and Engineering*, *495*.

Shen, L., Er, M. J., & Yin, Q. (2020). The classification for high-dimension low-sample size data. doi: https://doi.org/10.48550/arXiv.2006.13018

Tan, A. C., & Gilbert, D. (2003). Ensemble machine learning on gene expression data for cancer classification. Retrieved from `http://bura.brunel.ac.uk/handle/2438/3013`

Tiwary, S., Morales, J., & Kwiatkowski, S. e. a. (2018). "metastatic brain tumors disrupt the blood-brain barrier and alter lipid metabolism by inhibiting expression of the endothelial cell fatty acid transporter". *Sci Rep 8*, *8267*. doi: https://doi.org/10.1038/s41598-018-26636-6