Machine Learning in Cancer Detection

by

Shiyue(Vanessa) Wang

Tao(Noah) Zou

A Interactive Qualifying Project

Submitted to the Faculty

of the

WORCESTER POLYTECHNIC INSTITUTE

by

Jun 2022

APPROVED:
Smith, Therese

Chapter 1

Introduction and Related Work

Learning is a biological ability, and the current development of artificial intelligence has also enabled computers to have this ability. With the overall development of computer technology, the application of machine learning in public health is also becoming more and more extensive, and breakthroughs have been made in many fields such as medical imaging, clinical decision support, speech recognition, drug mining, health management, pathology, etc., which are also of great help to improve the diagnostic accuracy, safety and reliability of the medical system.

In public health-related research, there are a large number of samples - cases, such as influenza, tuberculosis, AIDS and other epidemics and infectious diseases case database, etc., which can not only constitute a text information set but also build a picture library of pathological features, which makes machine learning have the most basic conditions in public health research, and with the continuous diagnosis of patient diseases, data samples will continue to accumulate and enrich, which in turn can improve the "performance" of machine learning.

In Dmitrii Bychkov's[1] "Deep learning identifies morphological features in breast cancer predictive of cancer ERBB2 status and trastuzumab treatment efficacy", it uses

[1] Dmitrii Bychkov et al. 2021

machine learning to train and classify the status of the ERBB2 gene on the sample test set to predict the decision and treatment information of breast cancer. Since the disease is too complex, especially with many edge cases, the use of machine learning can help to better predict the direction of the disease and carry out targeted therapy. Additionally, the CNN trained with machine learning in the current study not only significantly and independently predicts ERBB2 status factors, but also classifies patient populations that benefit more from associated treatment options while identifying CISH ERBB2-positive cancer patients. Also, machine learning is used to identify edge case patients who were CISH ERBB2- negative but exhibited ERBB2-positive cancer-like according to CNN and had an unfavorable outcome (Dimitrii Bychkov et al., 2021).

According to scholars such as Rasool Fakoor[2], for tumor cells on the same microarray platform, machine learning can be applied to different cancers on the same platform by analyzing a single sample of a specific tumor. (For the feature learning that forms the basis for prostate cancer classification we can use samples from breast cancer, lung cancer, and many other cancers which are available in that platform.) The primary means of diagnosing patients by human doctors is much the same as machine learning (by analyzing samples to derive probabilities) (Rasool Fakoor et al., 2013). But their main difference is the size of the samples. Human doctors rely on their own clinical experience and patient samples to draw conclusions, and clinical experience is very limited (limited by time and space). But machine learning can refer to all the documented samples of known human cancers to compare with the obtained genetic samples.

[2] Rasool Fakoor et al. 2013

Chapter 2

Research Problems and Solutions

2.0.1 Research Problems

There are still many problems in the field of disease research using machine learning. The data problems include but are not limited to: lack of complex genotype phenotype association studies, highly sparse data, small sample size, etc., which will affect the judgment of the diseases[3], resulting in underfit or overfit issues to the training samples.

Also, according to scholars such as Konstantina Kourou[4], the rationality of the experimental design, the collection of appropriate data samples, the accuracy and validation evaluation of the classification results, and the size of the samples may all lead to the bias of the training set and thus the classification bias (Konstantina Kourou et al., 2015). Due to the diversity of conditions, machine learning techniques should be improved to be suitable for different difficult situations, in other words, to increase the diversity of learning ability. Today, most predictions use only molecular and clinical information to predict cancer outcomes. The lack of feature sets prevents machines from better integrating into complex cases.

Machine learning not only has technical difficulties to solve, but also ethical and

[3] Yuchen Yuan et al. 2021
[4] Konstantina Kourou et al. 2015

legal challenges. We have grouped these issues into three broad categories. The first category is machine learning that inevitably requires the collection and utilization of personal information. For technology companies, the infringement issues involved in the algorithm, how to use it legally without infringing, and the disclosure of personal information need to have substantial constraints[5]. The second category is how to determine the loss caused by the error of the machine learning training process. Machine learning is applied to cancer detection by having computers trained on patient data to help doctors to make decisions. In this process, the behavioral deviation caused by the slight misunderstanding of the computer system will cause errors in the judgment results, which are unavoidable. For example, the property damage and body damage and even the mental damage suffered by the patient should be judged. The results of the patient's examination through this technology may be different from the real situation, (the location of the cancer or the authenticity of the cancer will be challenged by traditional medicine). The third category is the legal nature of machine learning and its opacity to the general public. The most important part of machine learning is its structure, but whatever the method is, an invention or a product for a company, public understanding of the structure is vague. Furthermore, for the general public, the lack of expertise and the reluctance of companies to disclose algorithms to protect trade secrets can make them mistrustful. Removing the suspicions of the public and making them willing to believe and use this technology will be a huge help in the iteration and upgrade of the method.

[5] Seumas Miller et al. 2019

2.0.2 Research Solutions

Although there are many problems, there are also many targeted solutions. For data related problems, DeepGene's technology achieves a performance improvement of above 24 percent of the accuracy compared with general classification methods[6]. It uses mutation frequency to filter out most of the irrelevant genes, and uses index sparsity reduction (ISR) to convert gene data into indices of non-zero elements to significantly suppress the effects of data sparsity (Yuchen Yuan et al., 2021). From the perspective of neural networks, we found that the DNN classifier used by DeepGene can not only classify edge data, but also extract high-level features for more accurate classification. Furthermore, selecting the most informative subset of features to train different feature sets and using an SVM that accepts enough parameters for classification can better aid in diversity studies[7]. From the perspective of society, establishing a public data warehouse to collect valid cancer data sets that have been diagnosed can also better address the problem of data shortages. With the rapid development of HTTs, including genomics, proteomics, and imaging techniques, new types of input parameters have been collected. We should allow as many parameters as possible to be predicted by integrating genomic, clinical, histological, imaging, demographic, epidemiological, and proteomic data or different combinations or types of these (Konstantina Kourou et al., 2015). Among them, there is a new technique: baby machines[8]. It will initially accept simple cases. It uses relevant solutions to determine the probability distribution of the function space, thus providing solutions to more difficult problems in the future. By

[6] Yuchen Yuan et al. 2021
[7] Konstantina Kourou et al. 2015
[8] Ray J Solomonoff et al. 2006

providing a more difficult problem, it will correspondingly update the probability distribution of the solidion. More difficult solutions could be calculated by constantly itemizing and recursion[9].

We have some possible solutions to the legal and ethical implications of machine learning. The most important thing is to clarify the property rights of the algorithm and make it protected by patents and intellectual property rights[10]. There should be relevant laws to restrict the possible leakage of personal information, and information holders have the right to know and the right to choose whether to share personal information and to understand the whereabouts and uses of such information. Finally, the losses caused by algorithm errors should be compensated by the corresponding companies and restrained and supervised by the relevant departments.

[9] Ray J Solomonoff. 2006
[10] Ying Tao 2018

Chapter 3

Approaches

3.0.1 Theoretical

After analyzing the current approaches, we have considered possible improvements and created our own approach. We will first introduce our theoretical approach and apply it to our simulation.

First of all, we choose images instead of direct data for two reasons. First, the judgment of images reduces the bias of data parsing. The acquisition of direct data may be convenient, but there are also problems. For example, the data extracted based on images or MRI may deviate from the actual situation. Using pictures to conduct our machine learning will help us understand the reliability and reliability of the data accuracy.

Second, in practical applications, identifying pictures or image information is a more direct method. Specifically, training the machine to directly recognize pictures eliminates the step of converting image information to digital information, which increases the accuracy of data and the efficiency of analysis to a certain extent.

We intend to train our machine on the tissue color, size, location, texture, smoothness, and color around the diseased area so that it can finally judge whether the tumor is malignant or not based on the image information.

For the first step, we have to perform a dataset analysis and classify based on specific features. For the filter, we choose to use KNN (K-nearest neighbors algorithm) to find the fittest K value for the first centralized filtering under the conditions of the equipment. It has a time complexity of O(n), high accuracy, and insensitivity to outliers, which is more suitable for filtering unnecessary information, but because the memory usage is too large, we choose to use it for as much data as possible in classification and regression. After filtering the information, we need to extract the edge case from normal training to facilitate analysis using Adaboost. It is an additive model, and each model is built based on the error rate of the previous model, but it pays too much attention to the wrong data and less attention to the correct ones. However, high-precision classification can be obtained after successive iterations, which is most suitable for outlier data such as edge cases.

After filtering the unnecessary data, we decide to choose the Naive Bayesian network for comparison and classification. We use specific features as prerequisites, such as what a tumor looks like under a certain condition. However, this algorithm cannot learn the interaction between features. In the case of many conditions, this is a kind of classification that can be used for each specific condition. In this way, we can not only avoid the shortcomings of Naive Bayesian's inability to draw inferences from one case, but also improve the speed of data classification.

Before classification, we intend to use the sigmoid function to obtain the distribution of features. As for its advantages, the amount of calculation is very small and the speed is very fast. After this, we can extract a certain bounding limitation as a

precondition for plug-in before classification.

In general, the first choice would be logistic regression. If it is not pleasant, then its results can be used as a reference to compare and merge with other algorithms on the basis. Then, we will try a decision tree (random forest, for example) to see if we can drastically improve our model performance. Even if we don't use it as the final model, we can use the random forest algorithm to remove the noise variables in order to make the next feature selection. If the number of features and observation samples is particularly large, then using SVM may be an option when resources and time are sufficient. But we will consider the size of sample data and feature data to decide whether to choose SVM or not. High accuracy provides a good theoretical guarantee to avoid over-fitting, and even if the data is linearly inseparable in the original feature space, as long as a suitable kernel function is given, it is still a good choice to be used in classifying some parts of the data.

After the machine learning is completed, we will analyze the accuracy of our algorithms and models. We will find image datasets from the Internet to simulate our algorithm and compare the results of the simulation with the results of the dataset to obtain the correct rate. If the accuracy is below 80%, we will improve the algorithm and simulate it together again. Theoretically, we think that the success rate is 80%, which will be a practical, applicable model.

3.0.1 Experimental

In addition, we have compared several models of the whole training process. As

we are going to take in image features as the input, we choose to generate and compare the Receiver Operating Characteristic(ROC) curve based on either target tumor size or the colors in RGB scale.

Firstly, the decision tree (follow-up forest) approach is listed as a complete approach because it handles the interaction between features without stress and is non-parametric. Therefore, we don't have to be concerned about outliers or whether the data is linearly separableor not, but we will consider using logistic regression first with new data and new models. Because using logistic regression does not need to worry about the relationship between features.

Compared with decision trees and SVM machines, the accuracy rate of logistic regression is higher, and the model can be even easily updated with new data (using an gradient descent algorithm). If we need a probabilistic architecture (for example, to simply adjust the classification threshold, specify uncertainty, or obtain confidence intervals), or we want to quickly integrate more training data into the model later, we will use it. In order to prevent overfitting, we choose to select the appropriate model to use the decision tree. It is very important to choose an attribute to deal with the problem of the branch (pruning branch, for example). Here we may consider simple classification first. It is more suitable for dealing with samples with missing attributes and dealing with irrelevant features and can produce feasible and effective results for large data sources in a relatively short period of time. For the calculation that needs to find the optimal solution in the training process, we will first consider linear regression, because it is simple to implement and simple to calculate. The gradient descent can be used every time the optimal solution is to be calculated, and the cost is relatively small.

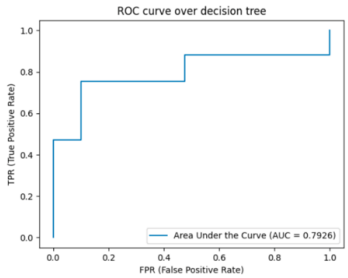

Image1-1: ROC curve of decision tree over RGB

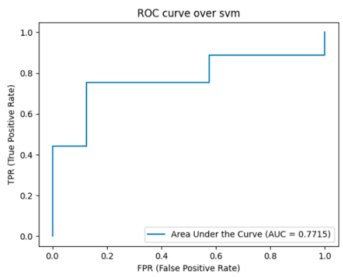

Image2-1: ROC curve of support vector machine over RGB

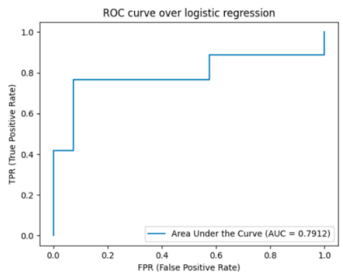

Image3-1: ROC curve of logistics regression over RGB

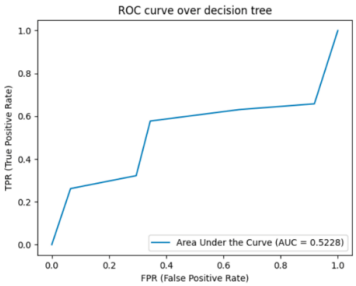

Image1-2: ROC curve of decision tree over size

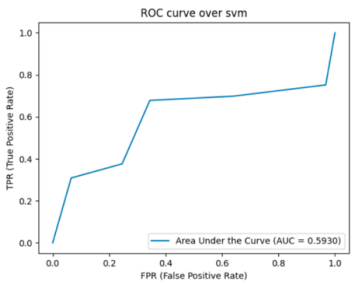

Image2-2: ROC curve of support vector machine over size

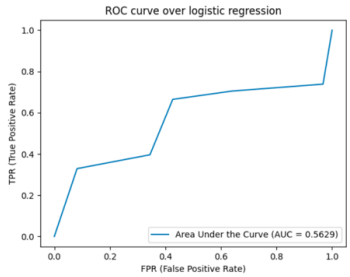

Image3-1: ROC curve of logistics regression over size

As we can see from Figure 1, it is easy to calculate that the average area under the curve(AUC) is 0.6577. For Figure 2, we can simply tell that the average AUC is nearly 0.6823. For Figure3, we got the average AUC to be nearly 0.6771. In short, the performance of SVM is the best. Thus, we decided to extend the performance tests of SVM by changing the kernel function.

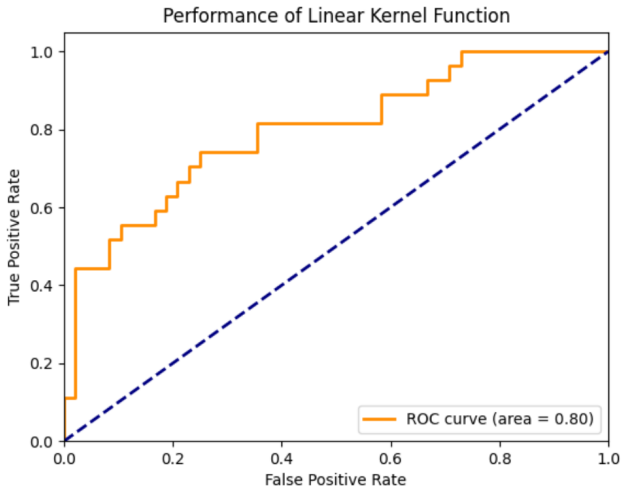

Figure 4-1: choosing linear regression as kernel function

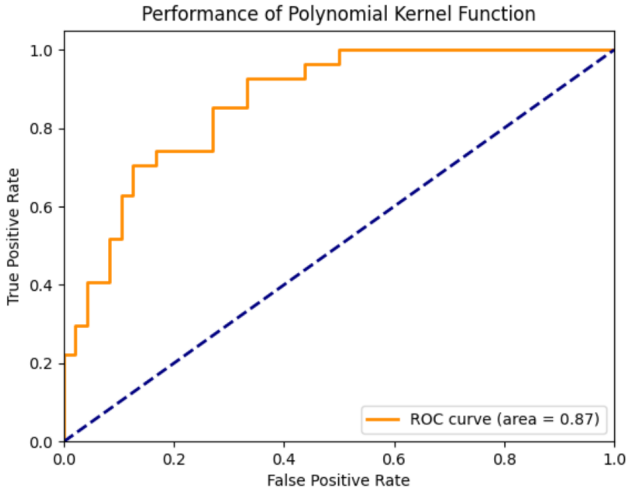

Figure 4-2: choosing polynomial regression as kernel function

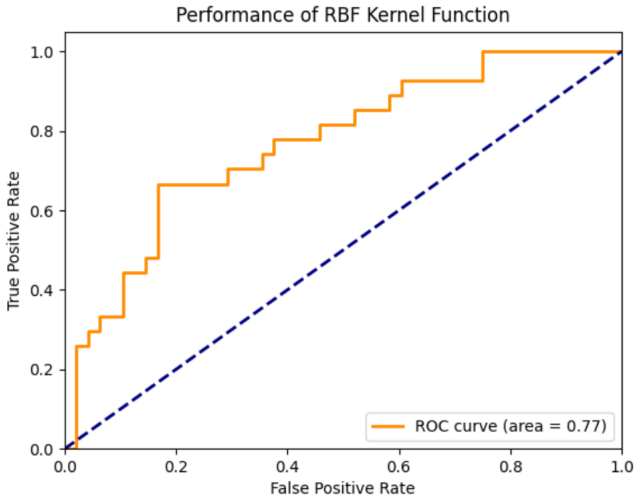

Figure 4-3: choosing Radial Basis Function as kernel function

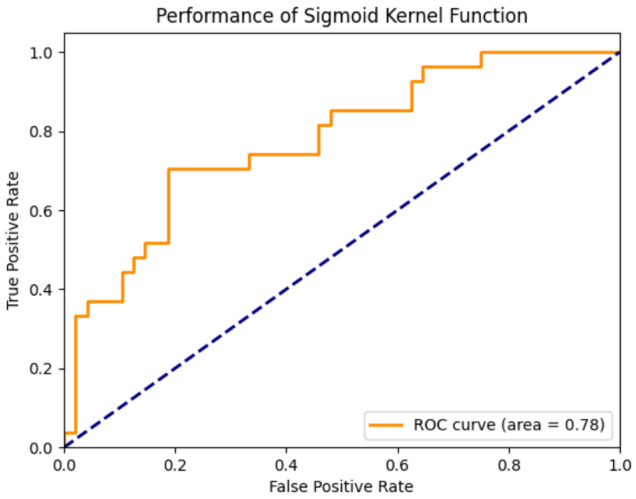

Figure 4-4: choosing Sigmoid Regression as kernel function

3.0.2 Conclusion

In summary, since the testing data for our comparison cases is relatively small and is manually refined, the test for the model is likely to produce overfitting or underfitting, resulting in inaccurate predictions. However, after comparing the four different kernel functions, we find that third degree polynomial regression has the highest score of performance. Thus, we made the preliminary decision to choose polynomial regression in the support vector machine for our general classification. The above performance test predicts the progression of three models in terms of classification and the final performance of SVM, but it is still a situation to be considered if the accuracy of the prediction will meet the future expectations or datasets.

Reference

[1] Dmitrii Bychkov et al. “Deep learning identifies morphological features in breast cancer predictive of cancer ERBB2 status and trastuzumab treatment efficacy”. In: Scientific reports 11.1 (2021), pp. 1–10.

[2] Rasool Fakoor et al. “Using deep learning to enhance cancer diagnosis and classification”. In: Proceedings of the international conference on machine learning. Vol. 28. ACM, New York, USA. 2013, pp. 3937–3949.

[3] Konstantina Kourou et al. “Machine learning applications in cancer prognosis and prediction”. In: Computational and structural biotechnology journal 13 (2015), pp. 8–17.

[4] Seumas Miller. “Machine learning, ethics and law”. In: Australasian Journal of Information Systems 23 (2019).

[5] Ray J Solomonoff et al. “Machine learning-past and future”. In: Dartmouth, NH, July (2006).

[6] Yuchen Yuan et al. “DeepGene: an advanced cancer type classifier based on deep learning and somatic point mutations”. In: BMC bioinformatics 17.17 (2016), pp. 243–256.

[7] Ying Tao. “Law Supervision of Machine Learning by Ying Tao 2018” In: Law Journal (2018).