Research Article

Arash Hooshmand*

Accurate diagnosis of prostate cancer using logistic regression

https://doi.org/10.1515/med-2021-0238
received September 2, 2020; accepted January 29, 2021

Abstract: A new logistic regression-based method to distinguish between cancerous and noncancerous RNA genomic data is developed and tested with 100% precision on 595 healthy and cancerous prostate samples. A logistic regression system is developed and trained using whole-exome sequencing data at a high-level, i.e., normalized quantification of RNAs obtained from 495 prostate cancer samples from The Cancer Genome Atlas and 100 healthy samples from the Genotype-Tissue Expression project. We could show that both sensitivity and specificity of the method in the classification of cancerous and noncancerous cells are perfectly 100%.

Keywords: machine learning, prostate cancer, diagnosis, transcriptome, RNA sequencing, high throughput technologies, logistic regression, classification

1 Introduction

Prostate cancer is one of the severe cancers in men. According to the US cancer statistics report for 2020, there are estimated 191,930 new cases of prostate cancer and 33,330 deaths because of it, and the importance of early diagnosis has repeatedly been emphasized [1]. Biologists have discovered many genes that are involved in specific cancers; for example, BRCA1 in breast cancer [2] and STAT3 in prostate cancer [3]. In diagnosis and cancer identification, histological examination is used as gold standard but it is a slow process and needs technical experts and suffers from large amount of variations among observers. In recent years, thanks to high throughput Omics technologies, we are no longer missing data but need novel methods and techniques to handle and analyze them; thus, bioinformatics and computers have found a solid ground to contribute in life sciences. One of the most applicable approaches to benefit from computer science in physiology and medicine is utilization of artificial intelligence to extract knowledge by computers out of big data generated by Omics technologies [4]. In this work, we have developed a logistic regression (LGR) system using general new generation of RNA Seq. data that can detect any prostate cancer, and hence will decrease the risk of mortality by correct diagnosis. The Omics technologies and their corresponding big data analysis tools are developing fast and getting cheaper and more widespread all the time. Currently, the third generation of sequencing methods such as quantum sequencing [5], nanopore sequencing [6], and single-molecule real-time sequencing [7] are making it possible even today for the wealthy people to benefit from expensive analyses, and if the current trend in advancements continues, it will not be a long way left to have commonplace analytical tools and services in each hospital and city. The advantage of machine learning is that as it gets more and more samples, its training would be more matured and more robust; therefore, there is a hope that the 100% accuracy that is achieved by a modest amount of data can be stabilized in the future when many patients and healthy people samples are fed to the system.

Computational techniques and tools are rapidly opening their positions in medical and pharmaceutical sciences too [8]. Different methods have been developed and tested in the last few decades and have returned great results in different fields of medicine including but not limited to cancer identification [9]. In this work, we have come up with a novel approach of applying LGR for cancer detection that is effective and robust. Using our method, cancerous tissue can correctly be identified, thus providing an opportunity to be controlled on time. This approach also offers a new direction for disease diagnosis while providing a new method to predict traits based on genomic information.

* **Corresponding author: Arash Hooshmand,** Department of Biomedical Engineering and Health Systems, School of Engineering Sciences in Chemistry, Biotechnology and Health, KTH Royal Institute of Technology, 11428 Stockholm, Sweden, e-mail: hooshmand@kth.se

Open Access. © 2021 Arash Hooshmand, published by De Gruyter. This work is licensed under the Creative Commons Attribution 4.0 International License.

2 Methods

In this project, we have used LGR algorithm from Sci-Kit Learn on 495 samples from The Cancer Genome Atlas (TCGA) research network and 100 samples of the Genotype-Tissue Expression (GTEx) project portal and directly fed the genome data to the machine to do heavy statistical calculations on our high dimensional data. The different parts of the method are clarified below. We use all the available data at the time of accessing the databases and have not ignored any sample.

2.1 Binary LGR

The LGR is a group of statistical techniques that aim to test hypotheses or causal relationships when the dependent variable is nominal.

Despite its name, it is not an algorithm applied in regression problems, in which continuous values are dealt with, but it is a method for classification problems, in which a binary value, i.e., either 0 or 1 is obtained. For example, a classification problem is to identify if a given tumor is malignant or benign. With the LGR, the relationship between the dependent variable, i.e., the statement to be predicted, with one or more independent variables, i.e., the set of features available for the model is determined. To do this, it uses a logistic function that determines the probability of the dependent variable. As previously mentioned, what is sought in these problems is a classification, so the probability must be translated into binary values for which a threshold value is used. If the probability values were above the threshold value, the statement is true and *vice versa*. Generally, this value is 0.5, although it can be increased or decreased to manage the number of false positives or false negatives [10].

In supervised classification methods the input data, usually seen as p points, are viewed as a p-dimensional vector (an array or ordered list of p numbers). Then the classifiers are more or less based on similar criteria, e.g., in the Bayesian classifiers, the classifier looks for a hyper surface that maximizes the likelihood of drawing the sample, or in SVMs, it looks for a hyperplane that optimally separates the points of one class from the other, which eventually could have been previously projected to a higher dimensional space. The LGR is a generalization of logits to distinguish samples that belong to one of the two different classes; hence, it is usually called binary LGR.

2.2 Feature selection

There are wrong perceptions in the computer science community about life science data that have prevented potential achievements, for instance, one is about the number of features [11] such as "it is obviously impractical to select all of the genes because mass dimensions will increase the computation cost." As a result, researchers usually try to reduce the assumed computational costs allegedly brought about by highly redundant dimensions and select a subset of features, i.e., genes to reduce the number of features and dimensions [12]. A strength point of our work is that we gave all the data corresponding to the whole-exome sequencing as feature inputs to the logistic regressor at once and it returned almost perfect results quickly and precisely. We thought of 19,627 different genes not as too many features but as different pixels of a less than 141×141 pixel photo, in which there are correlated pixels too, and it was a very light task for the machine to analyze such a low-resolution image and it took only seconds to classify the cancerous and noncancerous cells 100% precisely.

2.3 Model settings and evaluation

We have used LGR classifier also known as Logit or MaxEnt classifier from Scikit-Learn 0.23.1 with its default settings. Model evaluation produces measures to approximate a classifier's reliability. To distinguish between cancerous and noncancerous cells, as it is a binary classification, we use accuracy, precision, specificity, sensitivity, f1 score, several averaging techniques, and receiver operating characteristic curve to evaluate the model. We, indeed, use Sci-kit Learn Metrics Classification Report that returns precision, recall, and f1 score for each of two classes. In binary classification, recall of the positive class is called "sensitivity," and recall of the negative class is "specificity." In what follows, the terms and derivations from confusion matrix such as accuracy, specificity, sensitivity, and f1 score are given to review and compare:

Condition positive (P): the number of real positive cases in the data
Condition negative (N): the number of real negative cases in the data
True positive (TP) or hit
True negative (TN) or correct rejection
False positive (FP), false alarm, or type I error
False negative (FN), miss, or type II error

Sensitivity, recall, hit rate, or true-positive rate (TPR):

$$\mathrm{TPR} = \mathrm{TP}/\mathrm{P} = \mathrm{TP}/(\mathrm{TP} + \mathrm{FN}) = 1 - \mathrm{FNR}. \quad (1)$$

Specificity, selectivity, or true-negative rate (TNR):

$$\mathrm{TNR} = \mathrm{TN}/\mathrm{N} = \mathrm{TN}/(\mathrm{TN} + \mathrm{FP}) = 1 - \mathrm{FPR}. \quad (2)$$

Precision or positive predictive value (PPV) is the ratio of the correctly labeled samples by our program to all labeled ones in reality.

$$\mathrm{PPV} = \mathrm{TP}/(\mathrm{TP} + \mathrm{FP}) = 1 - \mathrm{FDR}. \quad (3)$$

Precision can be calculated only for the positive class, i.e., class 1 that shows cancer or can be evaluated for each one of the two classes independently treating each class as it is the positive class at time, and the latter is done in Sci-kit Learn Metrics Classification Report as shown in Table 1.

Negative predictive value (NPV):

$$\mathrm{NPV} = \mathrm{TN}/(\mathrm{TN} + \mathrm{FN}) = 1 - \mathrm{FOR}. \quad (4)$$

Miss rate or false-negative rate (FNR):

$$\mathrm{FNR} = \mathrm{FN}/\mathrm{P} = \mathrm{FN}/(\mathrm{FN} + \mathrm{TP}) = 1 - \mathrm{TPR}. \quad (5)$$

Fall-out or false-positive rate (FPR):

$$\mathrm{FPR} = \mathrm{FP}/\mathrm{N} = \mathrm{FP}/(\mathrm{FP} + \mathrm{TN}) = 1 - \mathrm{TNR}. \quad (6)$$

False discovery rate (FDR):

$$\mathrm{FDR} = \mathrm{FP}/(\mathrm{FP} + \mathrm{TP}) = 1 - \mathrm{PPV}. \quad (7)$$

False omission rate (FOR):

$$\mathrm{FOR} = \mathrm{FN}/(\mathrm{FN} + \mathrm{TN}) = 1 - \mathrm{NPV}. \quad (8)$$

Accuracy (ACC):

$$\begin{aligned}\mathrm{ACC} &= (\mathrm{TP} + \mathrm{TN})/(\mathrm{T} + \mathrm{N}) \\ &= (\mathrm{TP} + \mathrm{TN})/(\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}).\end{aligned} \quad (9)$$

The harmonic mean of precision and sensitivity or f1 score (F1):

$$\begin{aligned}\mathrm{F1} &= 2\cdot\mathrm{PPV}\cdot\mathrm{TPR}/(\mathrm{PPV} + \mathrm{TPR}) \\ &= 2\cdot\mathrm{TP}/(2\cdot\mathrm{TP} + \mathrm{FP} + \mathrm{FN}).\end{aligned} \quad (10)$$

Table 1: Classification report

Summary	Precision	Recall	f1 score	Support
Class 0	1.00	1.00	1.00	9
Class 1	1.00	1.00	1.00	51
Micro avg.	1.00	1.00	1.00	60
Macro avg.	1.00	1.00	1.00	60
Weighted avg.	1.00	1.00	1.00	60

As we are using Sci-kit Learn Metrics Classification Report to show the results as shown in Table 1, we also describe the meaning of micro avg, macro avg, and weighted avg. used in the report: Micro-average of precision (MIAP):

$$\mathrm{MIAP} = (\mathrm{TP1} + \mathrm{TP2})/(\mathrm{TP1} + \mathrm{TP2} + \mathrm{FP1} + \mathrm{FP2}). \quad (11)$$

Micro-average of recall (MIAR):

$$\mathrm{MIAR} = (\mathrm{TP1} + \mathrm{TP2})/(\mathrm{TP1} + \mathrm{TP2} + \mathrm{FN1} + \mathrm{FN2}). \quad (12)$$

Micro-average of f-score (MIAF) would be the harmonic mean of the two numbers above.

$$\mathrm{MIAF} = 2\cdot\mathrm{MIAP}\cdot\mathrm{MIAR}/(\mathrm{MIAP} + \mathrm{MIAR}). \quad (13)$$

Macro-average of precision (MAAP):

$$\mathrm{MAAP} = (\mathrm{Precision}\ 1 + \mathrm{Precision}\ 2)/2. \quad (14)$$

Macro-average of recall (MAAR):

$$\mathrm{MAAR} = (\mathrm{Recall}\ 1 + \mathrm{Recall}\ 2)/2. \quad (15)$$

Macro-average of f-score (MAAF) would be the harmonic mean of the two numbers above.

$$\mathrm{MAAF} = 2\cdot\mathrm{MAAP}\cdot\mathrm{MAAR}/(\mathrm{MAAP} + \mathrm{MAAR}). \quad (16)$$

Macro-average method is suitable to know how the system performs overall across different sets of data but should not be considered in any specific decision-making because it calculates metrics for each label and finds their unweighted mean, i.e., it does not take label imbalance into account, while in our case, the labels are highly imbalanced, i.e., 495 vs 100. On the contrary, micro-average is a useful tool and returns measures for decision-making especially when datasets vary in size because it calculates metrics globally by counting the total true-positives, false-negatives, and false-positives. Finally, weighted-average, according to Sci-kit Learn documentation on f1-score metrics, calculates metrics for each label and finds their average weighted by support (the number of true instances for each label). This alters "macro" to account for label imbalance; consequently, it can result in an F-score that is not between precision and recall.

3 Results

Genomic variation files of 595 samples including healthy people (100 individuals) and cancer patients (495 individuals) were obtained from the GTEx Project and the TCGA online database. The binary classification results of cancerous and noncancerous samples were great because

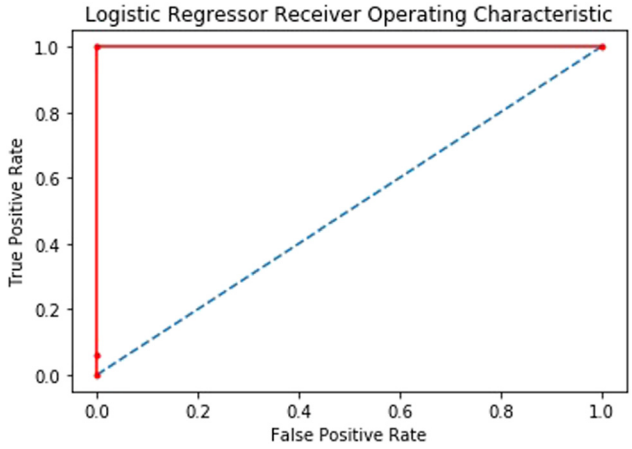

Figure 1: ROC curve of LGR classifier performance in distinguishing cancerous and noncancerous prostate cells.

the system can detect all cancerous and noncancerous samples correctly and as seen in the classification report shown in Table 1, the performance of the classifier is perfect with accuracy and precision of 100% and sensitivity and specificity of 1. In this classification, not only the accuracy is 100% but also the receiver operating characteristic's area under curve (ROC AUC) from prediction scores also would be 1 as seen in Figure 1.

4 Discussion

The classifier did its task perfectly with no error, at least on our available data. There are yet some aspects to reflect on. Although most TCGA prostate cancer (PRAD) comprise white men's samples, they have considered human variations to contain samples of different races and groups as well to represent the US demographic information fairly. As our method classifies all cancerous and non-cancer samples correctly using the information available in genomic variation, it can mean that the genetic signatures of cancer are detected universally without the need to consider racial or sexual differences.

Our work provided a new approach in application of computers using medical data that resulted in excellent classification between cancerous and noncancerous cells of the prostate. In this work, we did not reduce the dimension of input data and left all the statistical analysis to the computer, and it could do its job very well and distinguished the cancerous samples from healthy cells almost perfectly. We even did not need to balance the number of samples of each class and it shows that the difference between two classes is so much that providing hundreds of samples enables the machine to distinguish between two categories containing 495 and 100 samples perfectly. It is also useful to consider the fact that TCGA and GTEx data are not perfect and there are several rows of missing data for some of gene quantities in some samples, yet the data provided by these two projects are fairly clean and reliable and it was enough for our classifier to be able to do its classification 100% correctly. This system is trained now to receive any new person's RNA-seq data and recognize if the patient's prostate is cancerous or not. The limitation of our model is that it needs future collaboration with both hospitals and well-equipped laboratories and also needs the whole genome data of samples from the organ, and the involving labs should follow the same protocols to obtain the transcriptomics data. Therefore, we cannot add training data from other sources and databases to include as many samples as we want. Fortunately, we do not need to do it because our data have been enough to train the system and achieve perfect classification ability. Furthermore, an advantage of our approach is that we have used a classic interpretable method that is based on statistics, unlike other works such as Sun et al. [13] who have used complex neural networks that act as a black box and are not interpretable. Nevertheless, obtaining the whole-exome sequencing data of 19,627 genes as done by GTEx and TCGA on samples obtained from people's prostates is at research level and is not yet a cheap procedure or common practice for general hospitals. However, the New Generation RNA-seq protocols followed by GTEx and TCGA are well known and standard, and as technologies are developed rapidly, they are continuously getting cheaper and more practical than before. Meanwhile, the next topic of research can be finding suitable biomarkers in the blood that can detect healthy people and patients only by their blood tests.

Acknowledgments: The author thanks Houshmand family and their companies, especially Mr. Eng. GholamAbbas Houshmand, then Atash Houshmand, Shahab Houshmand, Shahin Houshmand, and Shadab Houshmand who financed all the study and research during several years. The author also thanks the library of KTH Royal Institute of Technology for financial support of the publication fee.

Ethical approval: This project does not need any extra personal/patient consent approval either because the data are normalized and do not reveal any private information, and whatever necessary with respect to the law is observed by the institutes publishing them.

Conflict of interest: The author is the only author of this article who has submitted it to De Gruyter's Open Medicine journal, and hence reiterates the consent to publish it in this journal. There are no competing interests and there is no need for any other consent approvals.

Data availability statement: The datasets generated during and/or analyzed during the current study are available in the GTEx and TCGA repositories that are publicly accessible on www.gtexportal.org and https://portal.gdc.cancer.gov.

References

[1] Siegel RL, Miller KD, Ahmedin J. Cancer statistics, 2020. CA Cancer J Clin. 2020;70(1):7–30.

[2] Atalay A, Crook T, Ozturk M, Yulug IG. Identification of genes induced by BRCA1 in breast cancer cells. Biochem Biophys Res Commun. 2002;299(5):839–46.

[3] Cocchiola R, Rubini E, Altieri F, Chichiarelli S, Paglia G, Romaniello D, et al. STAT3 post-translational modifications drive cellular signaling pathways in prostate cancer cells. Int J Mol Sci. 2019;20(8):1815.

[4] Nik-Zainal S, Memari Y, Davies HR. Holistic cancer genome profiling for every patient. Swiss Med Wkly. 2020;150:w20158.

[5] Di Ventra M, Taniguchi M. Decoding DNA, RNA and peptides with quantum tunnelling. Nat Nanotechnol. 2016;11(2):117–26.

[6] Branton D, Deamer DW, Marziali A, Bayley H, Benner SA, Butler T, et al. The potential and challenges of nanopore sequencing. Nat Biotechnol. 2008;26(10):1146–53.

[7] Thompson JF, Milos PM. The properties and applications of single-molecule DNA sequencing. Genome Biol. 2011;12(2):217.

[8] Goldenberg SL, Nir G, Salcudean S. A new era: artificial intelligence and machine learning in prostate cancer. Nat Rev Urol. 2019;16(7):391–403.

[9] Vamathevan J, Clark D, Czodrowski P, Dunham I, Ferran E, Lee G, et al. Applications of machine learning in drug discovery and development. Nat Rev Drug Discov. 2019;18(6):463–77. doi: 10.1038/s41573-019-0024-5.

[10] Cox DR. The regression analysis of binary sequences. J R Stat Soc Ser B Methodol. 1958;20(2):215–32.

[11] Pes B. Ensemble feature selection for high-dimensional data: a stability analysis across multiple domains. Neural Comput Appl. 2019;32(10):5951–73.

[12] Abeel T, Helleputte T, Van de Peer Y, Dupont P, Saeys Y. Robust biomarker identification for cancer diagnosis with ensemble feature se-lection methods. Bioinformatics. 2010;26(3):392–8.

[13] Sun Y, Zhu S, Ma K, Liu W, Yue Y, Hu G, et al. Identification of 12 cancer types through genome deep learning. Sci Rep. 2019;9(1):17256–9.