AUT Journal of Electrical Engineering

AUT J. Elec. Eng., 53(1) (2021) 41-46
DOI: 10.22060/eej.2020.18965.5375

Penalized Logistic Regression Models for Phenotype Prediction Based on Single Nucleotide Polymorphisms

S. R. Hosseini, F. Ghassemi* , M.H.Moradi

Department of Biomedical Engineering, Amir Kabir University of Technology, Iran

ABSTRACT: Most of the studies on phenotype differences, including some diseases, are based on studying some specific positions in the genome called Single Nucleotide Polymorphism (SNP). Some SNPs alone and some by interacting with others, play an important role in any phenotype or specific disease. Various models, including the regression models, are designed and implemented for the prediction of these diseases. In this paper, three penalized logistic models including Ridge, Lasso and Elastic Net (EN), are used to predict the risk of a specific disease, while overcoming the limitation of the classic logistic regression on high-dimensional SNP datasets. The models are implemented on 10000 samples of the SNP datasets of OWKIN-Inserm Institute, which contains 18124 SNPs. Among these three, the Lasso model with minimizer lambda indicate higher accuracy (73.73%) and AUC (83.54%). The model is also less complex, since it eliminates less related features as much as possible and keeps only the most informative ones. Additionally, getting better results with Lasso indicates that multicollinearity is either not existed between variables or is low and can be neglected.

Review History:

Received: Sep. 17, 2020
Revised: Dec. 23, 2020
Accepted: Dec. 25, 2020
Available Online: Jun. 01, 2021

Keywords:

Complex diseases prediction
genotype-phenotype associations
SNP
regression
penalized logistic regression

1- INTRODUCTION

The purpose of Genome-Wide Association Studies (GWAS) is to identify the genetic variants causing different phenotypes, one of which are called SNPs (Single Nucleotide Polymorphisms), that are very similar to mutations [1]. A SNP is a DNA sequence variation occurring when a single nucleotide adenine (A), thymine (T), cytosine (C), or guanine (G) in the genome differs between members of a species or paired chromosomes in an individual. For instance, two sequenced DNA fragments from different individuals, AAGCCTA to AAGCTTA, are different in a single nucleotide. This illustrates that there is a SNP at this specific position, and the two possible nucleotide variations – C or T – are said to be the alleles for this specific position. Finding agent loci in the genome and the relationship between themselves, and the phenotypes are important as some of these genetic variants in a person's genome along with his special environment or special diet can cause the relevant phenotype. Phenotypes are divided into two groups: quantitative phenotypes such as lipid level, blood pressure, height, weight, BMI and binary phenotypes, which are referred to as disease phenotypes or complex diseases. Complex diseases are essential to predict, as they can be controlled or even prevented by determining special environmental conditions, like what Armstrong and Tyler showed in their study of 5 children with phenylketonuria under a restricted regimen [2]. Phenylketonuria (PKU), which is an inherited disorder caused by a single gene, decreases the metabolism of the amino acid phenylalanine. Accordingly, the level of the amino acid, which is obtained totally through diet, increases in the blood and can cause serious health problems like mental disorders, intellectual disabilities and seizures. In the mentioned study, five children with PKU, aged from seven-month to four-year were put on a low-phenylalanine or phenylalanine-free diet to see if it can decrease their mental problems. The results of the study have shown that the disorders caused by PKU have been prevented in children with a restricted diet, especially if initiated in early age. This illustrates that genetic disorders can be prevented or controlled under certain environmental conditions.

Besides the importance of the environment in this type of phenotype, which is the critical solution to control the disease infestation, the best SNP selection plays a vital role in the risk prediction of the disease. SNPs are the most important parts of a genetic region, responsible for genetic disorders, though not all of them are agent. However, recent studies have indicated that certain SNPs are in strong association with the relevant phenotype, including complex diseases, which are responsible for more than two-thirds of the deaths around the world. SNPs are more advantageous than other gene data such as microarray gene expressions, due to their stability, high frequency and being easier and faster to collect [3]. However, they are still hard to work with, because of their huge size and significantly large number (up to one million) compared to the

*Corresponding author's email: mailto:ghassemi@aut.ac.ir limited number of samples (few hundreds or thousands). In addition, complex diseases are caused by not only some certain SNPs independently, but also the interactions among lots of SNPs, which needs more complex models to detect. Different models are used to detect agent SNPs and interactions for complex phenotypes risk prediction, most of which are based on Machine Learning methods. One of the popular models is the regression model, which allows to consider all markers together and is able to detect and remove weak effect in the presence of stronger causal effects [4]. Depending on the type of phenotype, which can be quantitative or binary, linear or logistic model is respectively implemented[5].

Multiple logistic regression (MLR) models, also known as multiple-SNP analyses or polygenic models, involve all the relevant SNPs used in the phenotype prediction models as explanatory variables. Thus, due to the nature of multiple regression, these models will be able to recognize the correlation and coupling in the SNPs, and make distinctions among them. MLR models have been implemented in different SNP-based studies on diseases such as Lung Cancer [6], Parkinson's Disease [7], Obesity [8] or many other diseases. However, the results of these models are reliable only when the number of samples are at least ten times as large as the number of SNPs to prevent overfitting. For high dimensional data, with a much greater number of SNPs, penalized regression models can be a better choice, aiming at shrinking the coefficients near to zero (Ridge regression models)[9], exactly toward zero (Lasso regression models) [10], or some of them close to zero, and some exactly to zero (Elastic Net models). Many studies have implemented penalized models on high dimensional SNP datasets to select the most relevant SNPs as a feature selection method [11, 12], or totally to predict the disease risk independently or in combination with other machine learning methods [12-16]. In addition to the compression and dimension reduction, which is mandatory in high dimensional data, penalized models are capable of identifying pertinent predictors in grossly underdetermined problems. The computational speed of these regularized models is also impressive, which makes them often outperform their un-regularized counterparts [13].

In this study, three penalized logistic regression models (Ridge, Lasso, and Elastic Net) have been implemented on data, finding the best hyperparameters for each, to find the highest accuracy among them. All reported results are the mean of 10-fold cross-validation, performed to assess the predictive performance of the models. The acquainted results of these models in this study demonstrated stronger prediction power, compared to other machine learning methods, which were implemented by other participants of the challenge. They have implemented methods including SVM, MLP, and X Gradient Boosted Tree on the whole samples of the train data, and got Area Under Curve (AUC) values below 0.80. The results are published and available on https://github.com/scouvreur/DiseasePredictionDNA.

The paper continues in three sections as methods, dataset, and results. In the first section, logistic regression models, along with penalized models and the implementation and results of each are explained. Later, the dataset and some related limitations are described in detail. Finally, the result of implemented models is discussed and compared to the results of previous efforts on the data.

2- PATIENTS AND METHODS

Patients Dataset

This study has been implemented on the dataset of the challenge "Disease prediction based on DNA data" held in 2018 by the OWKIN-Inserm Institute (https://challengedata2.ens.fr/en/challenges)[1]. The dataset includes 18124 SNPs of 26500 samples, of which only 10000 random samples are used in this study, for the limitation on available processing power. The data comprises 18124 related SNPs, each of which is not determined to be related to the phenotype. For confidentiality reasons, neither the name of the phenotype nor the SNPs are given. Additionally, the SNPs are defined by their alleles, coded in "0"s and "1"s in two different columns, meaning that for each couple of columns, corresponding to a given SNP, the more frequent allele in the dataset for that specific SNP is coded as "0", and the less frequent allele is coded as "1". It should be noted that there is no reference genome specified in the data, so that the "more frequent" allele coded by "0" does not necessarily represent the "reference" allele. Moreover, the SNPs relative positions have been permuted, meaning that two columns of "01" and "10" are considered the same.

3- METHODS

Regression is a statistical measurement estimating the relationship between one dependent variable, called outcome, and one or more independent variables called predictors or features. The most common form of regression is linear regression, in which the relationship between outcome and predictors is defined with a linear function. In linear regression, the only continuous outcome is permitted and they cannot predict categorical or binary outcomes. In this case, the logistic regression is defined, modeling the probability of a certain category or class by using a logistic function. The probability of disease for any sample in these models can be generally written as (1):

$$p\left(y|x,\beta\right)=(p_n)^y(1-p_n)^{1-y} \tag{1}$$

P_n is the probability of being diseased defined as (2), in which y is the state of the person to have a special disease (y=1) or not (y=0), x_n is the sequence of SNPs of the person, and β is the estimated coefficients for each SNP.

$$p_n=p\left(y=1|x_n,\beta\right)=\frac{1}{1+e^{-\beta^T x_n}} \tag{2}$$

As the goal of the logistic regression, the best coefficients are estimated by maximizing the likelihood (3) or log-

[1] The dataset is not freely available. It was only available for the participants of the challenge, but may be attained upon request.

likelihood (4) functions.

$$l\left(\beta \mid x_n, y\right) = p\left(y|x_n, \beta\right) = \prod_{n=1}^{N}(p_n)^y(1-p_n)^{1-y} = \prod_{n=1}^{N}\frac{e^{y\beta^T x_n}}{1+e^{\beta^T x_n}} \tag{3}$$

$$ll\left(\beta\right) = \sum_{n=1}^{N} y\log(p_n) + \left(1-y\right) \log(1-p_n) = \sum_{n=1}^{N} y\beta^T x_n - \ln(1+e^{\beta^T x_n}) \tag{4}$$

In the case of high dimensional data, in which the number of variables is far more than samples and multicollinearity (correlations between predictor variables) is more probable, these models will not fit since they are not able to detect multicollinearity and also more likely to overfit. In order to overcome such problems, penalized models are introduced. In these models, a penalty is added to the log-likelihood; so that the coefficients are estimated considering the multicollinearity between variables, and because of the penalty on large fluctuations, overfitting is avoided. After adding the penalty, estimation would occur by minimizing (5)

$$l_p\left(\beta_0;\beta;\right) = -l\left(\beta_0;\beta\right) + J\left(\beta\right) \tag{5}$$

where $l\left(\beta_0;\beta\right)$, the likelihood of the coefficients (β_0 is the constant coefficient and β refers to the other coefficients), denotes the unrestricted log-likelihood function phrased in (4), λ is the regularization parameter controlling the amount of shrinkage, of which the optimum should be found for the specific problem, and J(.) is the penalty function on the coefficient parameter. The reason that β_0 is separated here from other coefficients is that the intercept is not penalized explicitly. The penalty function is determined based on the method, which should be properly selected depending on the problem. Three of the frequently used penalization methods are Ridge, Lasso and Elastic Net, all of which are implemented and discussed in the following.

RIDGE REGRESSION

Ridge regression can create a parsimonious model when the number of predictor variables in a set exceeds the number of observations, or when a data set has multicollinearity, with the help of L2-regularization parameter, added to the log-likelihood. L2-regularization adds an L2 penalty, which equals the square of the magnitude of coefficients. In this model, all coefficients are shrunk by the same factor (none are eliminated). The study [9] has shown how Ridge estimators are used in the logistic regression model to obtain more realistic estimates for the parameters and to improve the predictive value of the model. The penalized log-likelihood function to be minimized in Ridge regression is (6), where I is the number of variables (SNPs). The tuning parameter λ controls the strength of the penalty term. When λ = 0, Ridge regression equals the ordinary logistic regression. If λ = ∞, all coefficients are shrunk to zero. The ideal penalty is therefore somewhere in between 0 and ∞.

$$l_p\left(\beta_0;\beta;\right) = -l\left(\beta_0;\beta\right) + \sum_{j=1}^{I}\beta_j^2 \tag{6}$$

LASSO REGRESSION

Unlike the Ridge model, Lasso standing for Least Absolute Shrinkage and Selection Operator, performs L1-regularization, which limits the size of the coefficients by adding an L1 penalty equal to the absolute value of the magnitude of coefficients (7). This sometimes results in the elimination of some coefficients altogether, which can yield sparse models. This particular type of regression is well-suited for high dimensional data, in which dimension reduction is needed or some special variables should be selected; but does not work as well as Ridge models on data showing high levels of multicollinearity. The same as the Ridge regression λ is basically the amount of shrinkage, increasing from 0 to ∞ sets none to all of the coefficients to zero. The more the λ, the more the coefficients are eliminated.

$$l_p\left(\beta_0;\beta;\right) = -l\left(\beta_0;\beta\right) + \sum_{j=1}^{I}|\,\beta_j\,| \tag{7}$$

ELASTIC NET

Elastic Net (EN) is a combination of both Ridge and Lasso, since it adds both L1 and L2 regularizations to the log-likelihood with the ratio of α (8). This has been performed to extend the models against some limitations of Lasso and Ridge. In high dimensional datasets, keeping all variables does not make much progress, thus, eliminating some of the least important and related variables to the response is needed to work with the data. On the other hand, in datasets with I number of variables and N number of samples, if I > N, the Lasso can select at most N variables. In addition to, Lasso models fail to make grouped selections, which refers to the selection of a group of genes having a high correlation with each other. In fact, Lasso models are only able to select one variable from a group and ignore the others. Elastic Net models with both regularizations overcome these problems. The L1 part of the penalty generates a sparse model and the L2 part removes the limitation on the number of selected variables, encourages grouping effect, and stabilizes the L1 regularization path. Accordingly, it is expected to demonstrate better results than both previous models.

$$l_p(\beta_0;\beta;) = -l(\beta_0;\beta) + (1-\alpha)\sum_{j=1}^{I}\beta_j^2 + \alpha\sum_{j=1}^{I}|\beta_j| \quad (8)$$

Preprocessing

The coding of the data (Table 1-a) was changed because: 1) the prediction models are mostly based on the number of less frequent alleles that each SNP has, and 2) the same role of “01” alleles and “10” alleles in the importance of the SNP to the phenotype. In the new coding, each SNP is coded in [“0”, “1”, “2”] in only one column, based on the number of its minor alleles, whether none, only one, or both of its alleles are minor (Table 1-b).

4- RESULTS

Implementation of Ridge regression:

After finding the optimal lambda for each fold, the Ridge regression model was implemented on data. The optimal lambda in each fold is found using cross-validation and calculating the error according to the log of each lambda. The plot for one of the folds is presented in figure 1-a. Optimal lambda balances the accuracy and simplicity of the model. As figure 1-a shows, the MSE increases with lambda. There are two best lambdas for each model: first, the one that minimizes the prediction error and gives the most accurate model (indicated by the left dashed vertical line), and the second is the largest lambda that gives the error within one standard error of the smallest (indicated by the right dashed vertical line).

Using the minimizer lambda leads to a higher accuracy but may increase the complexity of the model. However, in Ridge regression, since no variable is removed, practically the complexity of the model will not considerably decrease by the increase of lambda. As shown in Table 2, the complexity with minimizer lambda is even less with optimal lambda. Thus, the better choice would be the model with minimizer lambda, which gives the highest accuracy and smallest MSE. The accuracy and Area Under Curve (AUC), reported in table 2, is the mean of a 10-fold cross-validation that the ROC curve of each is shown in Figure 1 (b and c).

Implementation of Lasso regression:

As well as the Ridge regression models, Lasso models are implemented with both minimizer and optimal lambdas, found by cross-validation. As shown in Figure 2-a, the number of selected features is decreased by increasing the lambda. ROC curves of each fold along with AUC is indicated in Figure 2 (b and c) for both minimizer and optimal lambda.

In Lasso models, as the lambda increases, more coefficients are shrunk to zero and more features are eliminated and thus, the complexity of the model will significantly decrease. So, in the case that the number of selected features is more important to be low, the largest permitted lambda that does not increase MSE that much, would be a better choice for building the model. As shown in Table 2, the mean accuracy and AUC of the Lasso models are higher with minimizer lambda, and since the runtime for models with optimal lambda does not differ that much (meaning that the complexity is not enough to cause trouble), the model with minimizer lambda is considered as a better choice.

Implementation of EN regression:

Since Elastic Net regression models are eliminating variables like Lasso, complexity should also be considered when choosing the best lambda for the model. Nevertheless, lambda is not the only parameter to be optimized. Alpha, as the mixing parameter between Ridge (α=0) and Lasso (α=1) should also be optimized along with lambda to find the best EN model. When implementing cross-validation, the optimal λ and α are found among all their possible couples, like ones by which the model gives the least MSE. This will be repeated on each fold of the data to find the best parameters. Finding the best alpha and lambda in one random fold is visualized in Figure 3 (a and b). Based on Figure 3-a, the best choice for alpha would be the line including darkest points, since the color refers to the RMSE. Determining the best alpha would determine the minimizer and the optimal lambda, as shown in Figure 3-b.

ROCs of all 10-folds for both models with minimizer lambda and optimal lambda, are shown in Figure 3 (c and d). As the results in Table 2 shows, the model with minimizer lambda is better to choose compared to the model with optimal lambda, since the mean runtime for models with $\lambda_{minimizer}$ is not much larger than the mean runtime for models with $\lambda_{optimal}$. Moreover, the mean number of selected variables in each does not differ much to make considerable variations in the complexity of the models. Thus, in the case of using Elastic Net regression on the data, it seems better to build the model with $\lambda_{minimizer}$ rather than $\lambda_{optimal}$.

Selected variables do not differ enough to make considerable variations in the complexity of the models. Therefore, in the case of using Elastic Net regression on the data, it seems better to build the model with $\lambda_{minimizer}$ rather than $\lambda_{optimal}$.

5- DISCUSSION

This study has been performed to compare the efficiency of three penalized regression methods on SNP data related to a specific disease (undisclosed because of confidentiality reasons). Each method is implemented by finding the best hyperparameters. Two values for lambda is determined as the best values in all three models: 1) minimizer lambda that minimizes the MSE of the model, and 2) optimal lambda that is the largest lambda having an error within one standard error of the smallest MSE. All models are implemented with both lambdas. Among them, the Ridge regression was found to be the least efficient and most complex, since it involved all variables, some of which may be completely irrelevant to the phenotype and thus increase the complexity of the model. EN regression presented better results with the less complex model since it reduced the number of model variables to approximately 9.4% of all variables. The highest accuracy and AUC were achieved by Lasso regression, which eliminates irrelevant variables as much as possible, and thus, makes the model less complex.

As shown in Table 3, comparing the results of the three regression models, it can be surely said that Lasso regression outperforms two other methods, giving higher accuracy and AUC, especially with minimizer lambda rather than the optimal lambda, with less time of implementation. The penalty, added in Lasso, reduces the degree of overfitting that occurs in the model. In addition, the higher accuracy and AUC in Lasso (rather than Ridge) indicates that the multicollinearity between variables is negligible, meaning that the probability of randomly eliminating a relevant independent variable, which may be a multicollinear variable, is very low.

The only published effort on this data with different models such as SVM, MLP, and X Gradient Boosted Tree is available on https://github.com/scouvreur/DiseasePredictionDNA, which has reported AUCs below 0.80 for all cases.

REFERENCES

[1] J. Panigrahi, B. S. P. Mishra, and S. R. Dash, "Disease Prediction on the Basis of SNPs," in Emerging Technologies in Data Mining and Information Security: Springer, 2019, pp. 635-643.

[2] M. D. Armstrong and F. H. Tyler, "Studies on phenylketonuria. I. Restricted phenylalanine intake in phenylketonuria," The Journal of clinical investigation, vol. 34, no. 4, pp. 565-580, 1955.

[3] M. Waddell, D. Page, and J. Shaughnessy Jr, "Predicting cancer susceptibility from single-nucleotide polymorphism data: a case study in multiple myeloma," in Proceedings of the 5th international workshop on Bioinformatics, 2005, pp. 21-28: ACM.

[4] K. L. Ayers and H. J. Cordell, "SNP selection in genome-wide and candidate gene studies via penalized logistic regression," Genetic epidemiology, vol. 34, no. 8, pp. 879-891, 2010.

[5] S. Banerjee, L. Zeng, H. Schunkert, and J. Söding, "Bayesian multiple logistic regression for case-control GWAS," PLoS genetics, vol. 14, no. 12, p. e1007856, 2018.

[6] J. L. Weissfeld et al., "Lung cancer risk prediction using common SNPs located in GWAS-identified susceptibility regions," Journal of Thoracic Oncology, vol. 10, no. 11, pp. 1538-1545, 2015.

[7] Z. Zhu, D. Yuan, D. Luo, X. Lu, and S. Huang, "Enrichment of minor alleles of common SNPs and improved risk prediction for Parkinson's disease," PloS one, vol. 10, no. 7, p. e0133421, 2015.

[8] C.-F. Hung et al., "A genetic risk score combining 32 SNPs is associated with body mass index and improves obesity prediction in people with major depressive disorder," BMC medicine, vol. 13, no. 1, p. 86, 2015.

[9] S. Le Cessie and J. C. Van Houwelingen, "Ridge estimators in logistic regression," Journal of the Royal Statistical Society: Series C (Applied Statistics), vol. 41, no. 1, pp. 191-201, 1992.

[10] R. Tibshirani, "Regression shrinkage and selection via the lasso," Journal of the Royal Statistical Society: Series B (Methodological), vol. 58, no. 1, pp. 267-288, 1996.

[11] T. T. Wu, Y. F. Chen, T. Hastie, E. Sobel, and K. Lange, "Genome-wide association analysis by lasso penalized logistic regression," Bioinformatics, vol. 25, no. 6, pp. 714-721, 2009.

[12] Z. Wei et al., "Large sample size, wide variant spectrum, and advanced machine-learning technique boost risk prediction for inflammatory bowel disease," The American Journal of Human Genetics, vol. 92, no. 6, pp. 1008-1012, 2013.

[13] S. Okser, T. Pahikkala, A. Airola, T. Salakoski, S. Ripatti, and T. Aittokallio, "Regularized machine learning in the genetic prediction of complex traits," PLoS genetics, vol. 10, no. 11, p. e1004754, 2014.

[14] G. Abraham, A. Kowalczyk, J. Zobel, and M. Inouye, "Performance and robustness of penalized and unpenalized methods for genetic prediction of complex human disease," Genetic epidemiology, vol. 37, no. 2, pp. 184-195, 2013.

[15] D. Shigemizu et al., "The construction of risk prediction models using GWAS data and its application to a type 2 diabetes prospective cohort," PLoS One, vol. 9, no. 3, p. e92549, 2014.

[16] S. Cherlin, R. A. Howey, and H. J. Cordell, "Using penalized regression to predict phenotype from SNP data," in BMC proceedings, 2018, vol. 12, no. 9, p. 38: BioMed Central.

[17] T. Minami, H. Nanto, and S. Takata, "Highly conductive and transparent aluminum doped zinc oxide thin films prepared by RF magnetron sputtering," Japanese Journal of Applied Physics, vol. 23, no. 5A, p. L280, 1984.

HOW TO CITE THIS ARTICLE

Hosseini, S.R., Ghassemi, F., Moradi, M.H. Penalized Logistic Regression Models for Phenotype Prediction Based on Single Nucleotide Polymorphisms. AUT J. Elec. Eng., 53(1) (2021) 41-46.

DOI: 10.22060/eej.2020.18965.5375