Iran J Sci Technol Trans Sci
https://doi.org/10.1007/s40995-022-01292-8

RESEARCH PAPER

A Method for Cancer Genomics Feature Selection Based on LASSO-RFE

Chen Ai[1]

Received: 22 January 2022 / Accepted: 4 April 2022

Abstract
A more efficient feature selection method was developed to screen genes corresponding to specific cancers to further investigate their pathogenesis. The LASSO-RFE model, a last absolute shrinkage and selection operator (LASSO) classifier based on the idea of recursive feature elimination (RFE), was constructed. To verify the efficiency of the proposed algorithm, performance tests were conducted by using four kinds of gene expression RNA sequences publicly available in The Cancer Genome Atlas (TCGA). The numerical experiments were used to illustrate that the proposed LASSO-RFE enables a higher accuracy of the classification prediction model and a clearer biological interpretability of the selected gene features compared with three typical feature selection algorithms. The experimental results showed that LASSO-RFE effectively reduced tens of thousands of features in the original data to three dimensions and provided better performance for the classification model than mutual information, L1-SVM and tree-based selection method. This model retains the ability of the common LASSO algorithm to filter and remove redundant and irrelevant features, and enhances the biological interpretability according to RFE, which was compared with the traditional feature reduction methods. In this paper, only a limited number of data cases have been validated, and the application of LASSO-RFE with more recent data remains to be further investigated.

Keywords LASSO · Recursive feature elimination · Feature selection · Cancer genome

1 Introduction

Gene microarray, as a technology used to measure tens of thousands of gene expression profile data simultaneously, can effectively assist the researchers in investigating the mechanism of pathogenesis, diagnosis of the disease, and drug development in the genetic dimension (Golub et al. 1999; Ramaswamy and Golub 2002). However, as gene expression profile data suffer from the small sample size, high dimensionality, positive and negative sample imbalance, and the most of the data in gene microarrays are usually redundant information. There are a large number of noisy genes, thus, using raw data for diagnosis is inefficient: both in terms of the speed of running algorithms and the interpretability of results. In order to better extract and exploit the valuable information hidden in these data and select representative genes with the best classification performance (Tinker et al. 2006), more superior performance feature selection methods have to be developed.

Domestic and foreign scholars have proposed many methods for genetic data feature selection to tackle this problem. Early scholars often used two methods (Molina et al. 2002): (1) Filter and (2) Wrapper. Filter methods for gene feature selection are independent of the classifier, and Wrapper methods for gene selection are related to specific classifiers, including support-vector machines (Guyon et al. 2006). Filter methods can handle large amounts of data simultaneously and are faster but require algorithms that are linearly separable for individual features. In contrast, the actual situation is often a complex combination of multiple genes to express certain biological information, so Filter methods often have significant errors in gene feature selection (Peng et al. 2010). The accuracy of the Wrapper method is better than the Filter method, but it is computationally intensive due to the need to repeat the training data, and it tends to fall into local optimal solutions (Ramaswamy and Golub 2002).

✉ Chen Ai
aichen@email.ncu.edu.cn

[1] School of Sciences, Nanchang University, Nanchang 330000, Jiangxi, China

Published online: 20 April 2022

Therefore, based on the characteristics of Wrapper methods, researchers began to develop more effective algorithms. SVM is widely used in gene feature selection due to its better processing of sparse small sample data with noise. Guyon et al. (2002) first proposed the support-vector machine recursive feature elimination (SVM-RFE) method (Guyon et al. 2002), which removes the least relevant features in order of importance and then repeats iterations in the remaining features until all the features have been traversed.

Another class of algorithms capable of feature selection is linear regression models in a regularization framework. Tibshirani et al. (1996) proposed the L1 parametric regularization LASSO (Tibshirani 1996), which can accomplish both the training process of the learner and the feature selection process and has many good properties. The basic idea of LASSO is to minimize the sum of squared residuals under the constraint that the sum of absolute values of regression coefficients is less than a constant, thus generating some regression coefficients close to or even equal to 0, thereby obtaining a more interpretable model. However, the determination of this constant lacks theoretical support, and in general, cross-validation and manual observation are required to find the most appropriate parameter values quantitatively.

The RFE idea is to use an iterative approach to find the optimal parameters precisely and automatically. Based on the above research background, this paper applies the RFE recursive idea to feature selection in the LASSO algorithm, taking into account the removal of redundant and irrelevant features from the sample data and the interpretability of the screening results in the biological sense and establishes the LASSO-RFE mathematical model. The experimental results show that compared with three other feature selection methods, the proposed LASSO-RFE model makes the corresponding classification algorithm more accurate, and precision and recall are also the highest in most cases. Meanwhile, it also deciphers the practical interpretation of biological meanings.

The rest organization of this paper is as follows: First, it briefly describes the theoretical basis of the LASSO algorithm and how the RFE idea is applied to the LASSO algorithm. Then, it gives the rubric, data sources, and corresponding results of each of the numerical experiments used for validation, in addition to images that facilitate visual determination of the classification.

2 Theoretical Analysis

2.1 LASSO Algorithm

The LASSO method proposed by Tibshirani et al. (1996) is derived from Non-negative Garrote algorithm (Breiman 1995), whose objective function can be summarized in the form of Eq. (1).

$$\sum_{i=1}^{n}\left(y_i-\alpha-\sum_j c_j\beta_j x_{ij}\right)^2 \tag{1}$$

The algorithm is based on the idea of least squares estimation, which compresses the estimated regression coefficients by controlling for nonnegative factors. Breiman (1995) conducted extensive numerical experiments to show that the Non-negative Garrote algorithm enhances the interpretability of the model due to the removal of a large number of variables close to zero but nonzero in the original data. However, since the algorithm's results depend on the size of the least-squares estimate, there are cases of overfitting or highly correlated variables, which in turn affect the prediction accuracy. The LASSO algorithm was modified to avoid this drawback.

It may be useful to let the number of features be p, the sample size be n, and the data set be $(X_i; y_i), i = 1, 2, ..., n$. where $X_i = (x_{i1}, x_{i2}, ..., x_{ip})$ and y_i are the set of features and observations, respectively. Usually the observations y_i are considered to be independent of each other and the feature x_{ij} are normalized, i.e., they satisfy $\sum_i x_{ij} = 0, \sum_i x_{ij}^2 = 1$. Let $\beta = (\beta_1, \beta_2, ..., \beta_p)^{\mathrm{T}}$. Then the estimate (α, β) of the LASSO algorithm can be defined as:

$$(\alpha,\beta)=\arg\min\left\{\sum_{i=1}^{n}\left(y_i-\alpha-\sum_j\beta_j x_{ij}\right)^2\right\}\ s.t.\ \sum_j|\beta_j|\le t \tag{2}$$

where t in Eq. (2) is a nonnegative parameter, which is a L1-parametric penalty function for the regression coefficient β_j. Thus, Eq. (2) can be abstracted from the objective function as $F(\beta)$, i.e., we have

$$F'(\beta)=F(\beta)+\lambda\beta_1 \tag{3}$$

Its corresponding gradient is:

$$\nabla F'(\beta)=\nabla F(\beta)+\lambda sign(\beta) \tag{4}$$

where $sign(\beta)$ in Eq. (4) is the sign function. Let $\beta^* = \arg\min\{F(\beta)\}$, then the quadratic approximate solution of the objective function under Taylor expansion can be expressed in the following form:

$$F'(\beta) = F(\beta^*) + \sum_i \left[\frac{1}{2} H_{i,i}(\beta - \beta^*)^2 + \lambda|\beta|\right] \tag{5}$$

where H denotes the Hessian matrix. Since β^* is the optimal solution of $F(\beta)$, i.e., $\nabla F(\beta^*) = 0$, the formula for finding the gradient for each regression coefficient can be expressed as:

$$\nabla_{\beta_i} F'(\beta) = H_{i,i}(\beta_i - \beta_i^*) + \lambda sign(\beta_i^*) = 0 \tag{6}$$

The solution yields:

$$\beta_i = \beta_i^* - \frac{\lambda}{H_{i,i}} \cdot sign(\beta_i^*) \tag{7}$$

Equation (7) can be equivalently expressed as:

$$\beta_i = sign(\beta_i^*) \max\left\{|\beta_i^*| - \frac{\lambda}{H_{i,i}}, 0\right\} \tag{8}$$

It is easy to see that the control of parameter t can make the regression coefficients smaller overall, whereby some regression coefficients will shrink $\lambda/H_{i,i}$ units toward 0, and some will even be directly equal to 0. In the case of larger data and fewer parameters, the parameter t is usually chosen with reference to the Bayesian information criterion (BIC), that is

$$\min_t \{BIC(t) = -2\ln(L) + k\ln(n)\} \tag{9}$$

where L is the likelihood function, k is the number of parameters, and n is the sample size.

2.2 LASSO-RFE Model

The choice of regression coefficient threshold is the core of LASSO for feature selection. Regarding the constraints in the LASSO algorithm, specific manually determined parameters may be easy to understand, but it does not guarantee the performance of the model. A stable feature selection method should be repeated for the training data, and for each feature elimination, a test about the impact of the feature on the model performance should be performed. According to this criterion, redundant or irrelevant features are eliminated one by one in order of decreasing impact, preserving as much valid information contained in the data as possible. However, this idea is not applied to LASSO as a whole but at each iteration step.

Moreover, as one type of Wrapper, RFE iteratively constructs the model, then selects the best or worst features based on the coefficients, removes the selected features, and then repeats the process on the remaining features until all the features have been traversed. The order of the eliminated features in this process is the ranking of the features.

LASSO-RFE takes LASSO as the underlying algorithm and improves the feature selection with the RFE idea, retaining the advantage that LASSO effectively removes invalid features while strengthening the interpretation of biological meaning, making the constructed features more representative. After irrelevant features are removed at one time, the next iteration continues to build the LASSO model and continues to evaluate among the remaining features until the parameter $t = 0$. Theoretically, noisy, redundant, or irrelevant features are removed one by one in this process. The specific implementation flow of the LASSO-RFE model is shown in Fig. 1.

In order to speed up the model, the model removes all the features when they have exactly equal absolute values of the regression coefficients. Although this approach has a theoretical potential for misclassification, it does not significantly affect classification prediction in practice because it occurs mainly at the beginning of dimensionality reduction, for example, when all features with regression coefficients strictly equal to zero are removed at once. It is a very small probability that the absolute value of the regression coefficients of the identified key features and an unrelated feature are strictly equal.

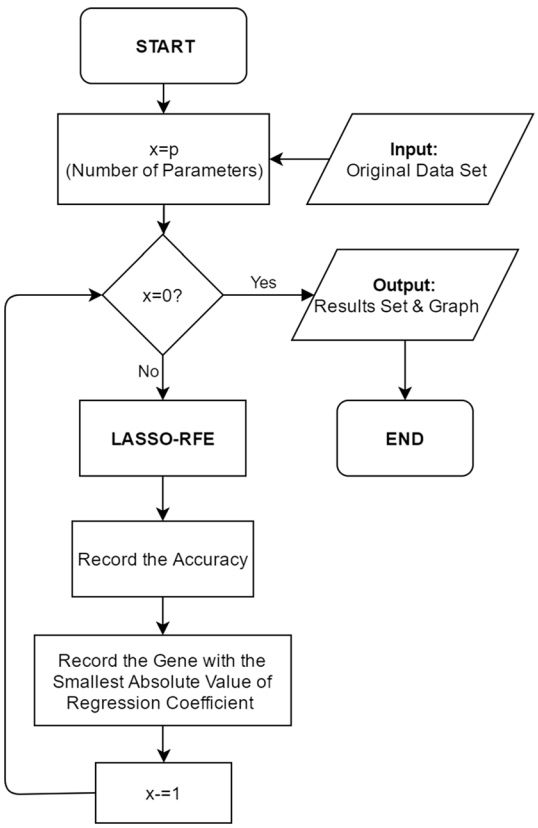

Fig. 1 The flowchart of LASSO-RFE

Table 1 Brief information about verification data

Data name	Number of genes	Number of samples	Number of diseased	Number of non-diseased
Breast cancer (BRCA)	60,483	1217	1118	99
Colon adenocarcinoma (COAD)	60,483	512	471	41
Head and neck cancer (HNSC)	60,483	546	502	44
Stomach cancer (STAD)	60,483	407	375	32

Table 2 Abbreviated representation of the HNSC data set

No	X_1	X_2	X_3	X_4	X_5	$\cdots$	X_{60481}	X_{60482}	X_{60483}	D
1	8.7879	0	11.0546	10.2467	1.5849	$\cdots$	7.6795	10.7927	11.9436	1
2	12.0647	2.8073	11.2929	9.9053	2.3219	$\cdots$	9.8564	9.7992	12.8362	1
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
315	11.7034	13.2849	10.1522	9.79441	5.3923	$\cdots$	10.5717	13.7613	11.5304	-1
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
545	10.7714	2.8073	10.4968	11.5328	1.5849	$\cdots$	8.2288	10.3609	10.6063	1
546	11.2118	2.3219	12.1215	11.3788	1.5849	$\cdots$	11.8301	12.2618	11.5157	1

3 Model Validation

3.1 Experimental Data and Evaluation Indexes

In this paper, a total of four common sets of publicly available mRNA-seq counts data from the TCGA database are used to validate the performance of the LASSO-RFE model (Zou et al. 2019), that is, the effect of classification for the data processed by it. The sequences data were converted to numerical data using FPKM calculation and the descriptive statistics are shown in Table 1.

To facilitate the observation of data types and data characteristics, the original data set is partially displayed as shown in Table 2, taking HNSC as an example. Where NO denotes the serial number, X_i denotes the code corresponding to the name of gene No i; $D = 1$ denotes as diseased and $D = -1$ denotes normal.

A reasonable means of performance evaluation can appropriately reflect the strengths and weaknesses of the model. The application context of this paper is a dichotomous classification problem, so the Confusion Matrix is used to show the base case of the classifier for prediction. Since the sample dependent variable is only 1 (sick) or -1 (normal), four scenarios are possible for the predicted and actual results of the model.

1. The actual outcome is sick and the predicted outcome is sick, noted as TP (True Positive).
2. The actual outcome is sick and the predicted outcome is normal, noted as FN (False Negative).

Table 3 Evaluation index and calculation formulas

Evaluation metrics	Equation
Accuracy	$A = (TP + TN)/(TP + TN + FP + FN)$
Precision	$P = TP/(TP + FP)$
Recall	$R = TP/(TP + FN)$
Robustness (F1)	$F1 = 2 \cdot P \cdot R/(P + R)$

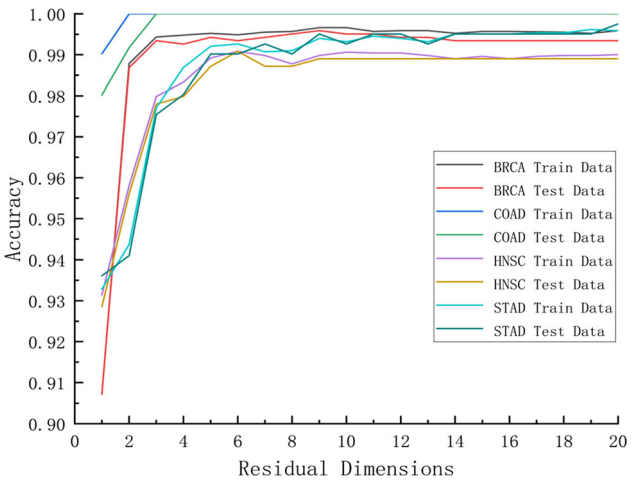

Fig. 2 The effect of the last 20 gene features on the prediction accuracy

3. The actual result is normal and the predicted result is sick, noted as FP (False Positive).

Table 4 Three principal genes filtered by LASSO-RFE

Data set	Antepenultimate feature	Penultimate feature	Last feature
BRCA	MFAP5	COL10A1	FGF1
COAD	MYOC	KRT80	OTOP2
HNSC	COL10A1	AIM2	ADIPOQ
STAD	CCN5	COL10A1	ESM1

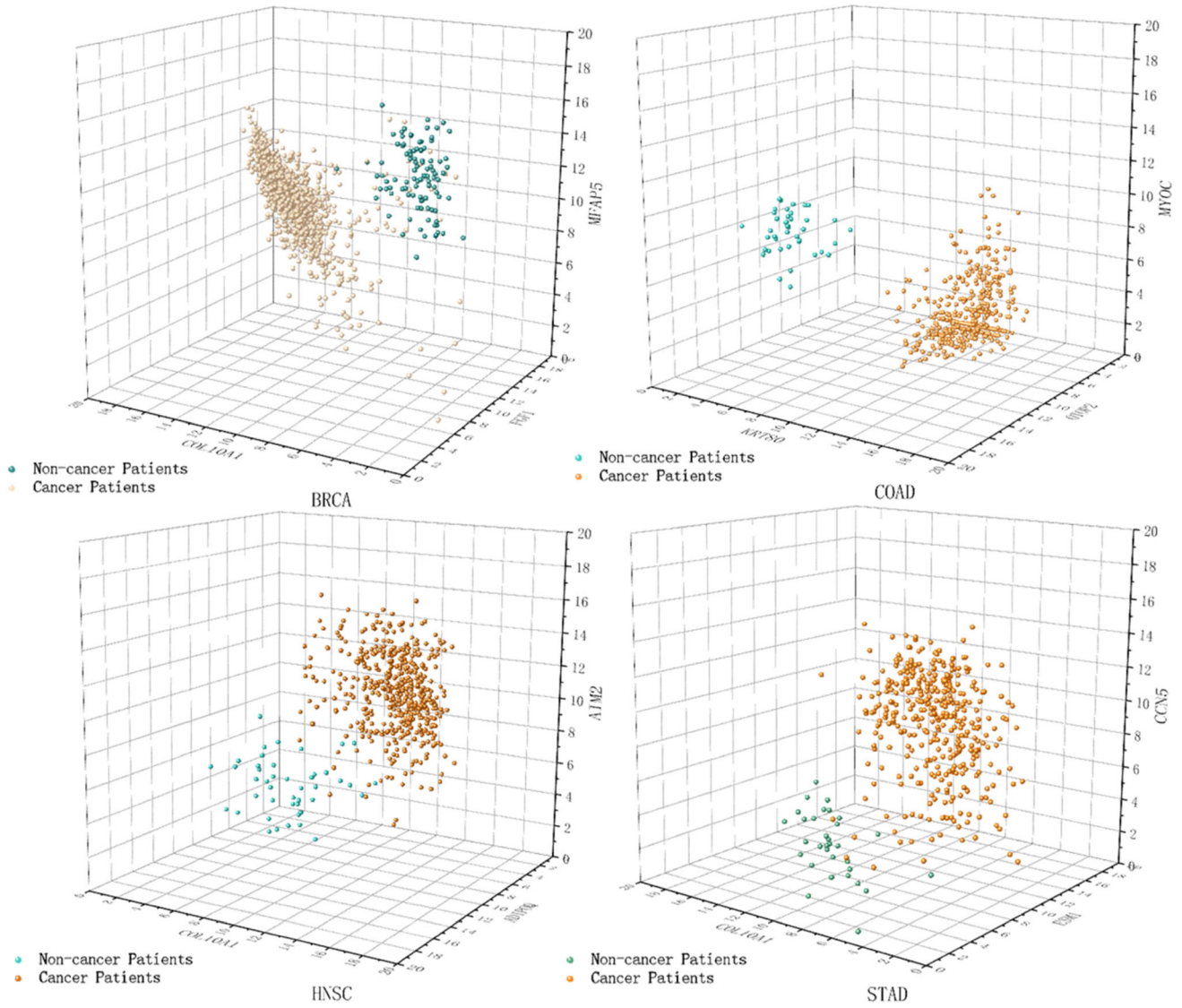

Fig. 3 Data distributions of features filtered by LASSO-RFE

4. The actual outcome is normal and the predicted outcome is normal, noted as TN (True Negative).
5. From this, a confusion matrix can be constructed and it generates some derived important indicators, such as Accuracy, Precision, Recall, and Robustness. Corresponding formulas are shown in Table 3.

3.2 Experimental Results

To minimize the randomness of the data for validating the model, a tenfold cross-validation was performed when the original data were processed according to the LASSO-RFE model, and the average of the prediction accuracies of the training and test sets were recorded separately as the results. Since most of the processes of dimensionality

Table 5 Detailed comparison of LASSO-RFE and other algorithms

Data Set	Indicators	RFE (%)	MI (%)	L1-SVM (%)	TBS (%)
BRCA	Accuracy	**98.76**	93.48	97.21	97.41
	Precision	99.18	95.82	98.36	**99.44**
	Recall	**97.67**	92.31	96.60	95.89
	F1	**98.42**	94.03	97.47	97.63
COAD	Accuracy	**100.00**	99.29	97.62	99.64
	Precision	**100.00**	**100.00**	98.08	**100.00**
	Recall	**100.00**	98.73	97.67	99.36
	F1	**100.00**	99.36	97.87	99.68
HNSC	Accuracy	**97.99**	88.31	83.19	96.88
	Precision	**98.20**	86.69	85.31	96.65
	Recall	97.61	93.43	84.46	**97.81**
	F1	**97.90**	89.93	84.89	97.23
STAD	Accuracy	**98.04**	90.05	72.90	97.13
	Precision	**100.00**	90.55	73.52	**100.00**
	Recall	**96.53**	92.00	80.80	94.93
	F1	**98.24**	91.27	76.12	97.40

Bold stands for the best situation among four algorithms

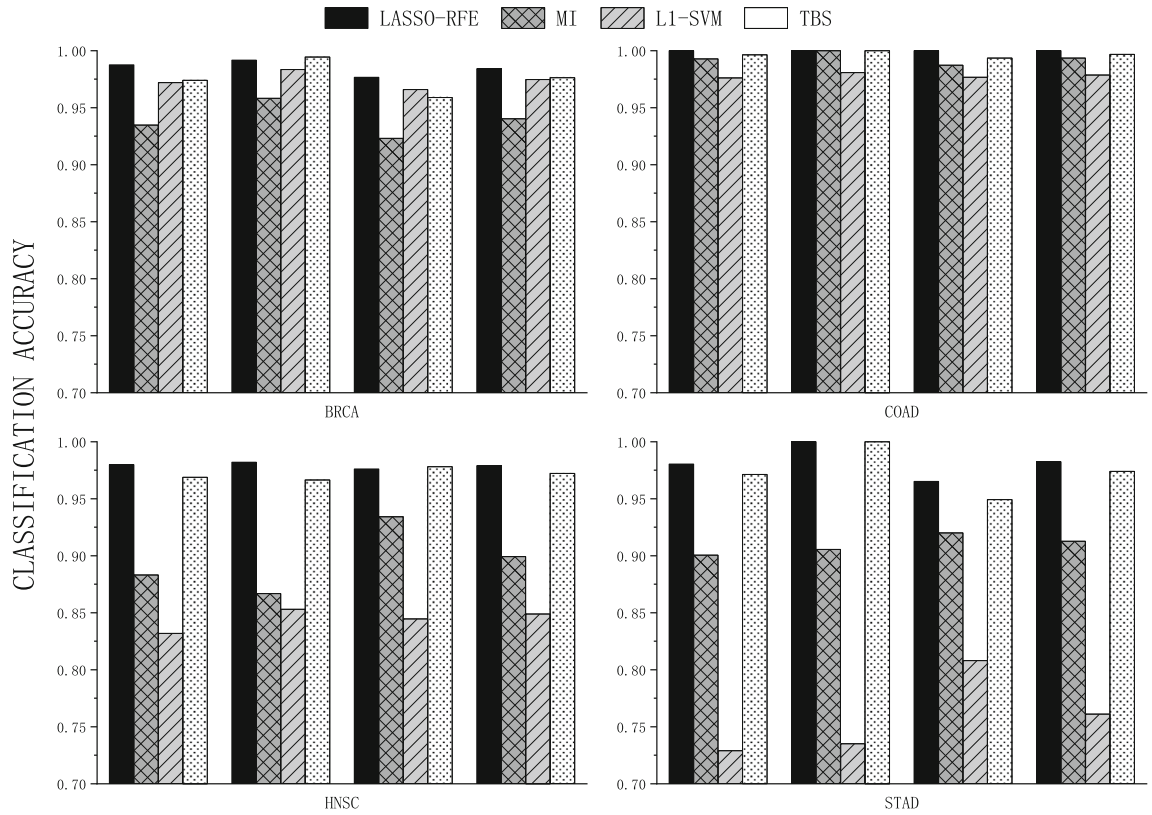

Fig. 4 Intuitive comparison of LASSO-RFE and other algorithms

reduction have little valid information that significantly affects the results, only the dimensionality reduction of the last 20 genetic features is shown in the main text, as shown in Fig. 2.

The feasibility of data visualization and the influence of dimensionality on prediction performance were considered comprehensively, and three dimensions were selected as the best dimension for feature selection in this paper. After

reducing the original number of 60,483 features to 3, it both significantly enhances the biological significance of the model and can well distinguish positive and negative samples, as shown in Table 4 and Fig. 3.

Figure 3 is the scatter diagram of data after feature selection, which intuitively shows that the remaining data is easy to classify. On the one hand, it indicates the high prediction performance given in Table 5, on the other hand, it shows the effectiveness of the model for feature selection.

Furthermore, in the limited datasets covered in this paper, the results in Table 4 show that there are 3 types of cancers that are strongly associated with the COL10A1 gene. In fact, COL10A1 as a tumor biomarker upregulated in a wide variety of tumors (Chapman et al. 2012), its expression does significantly affect the prognosis of multiple cancers (Huang et al. 2018; Chen et al. 2019). It has been supported by the published literature in recent years (Li et al. 2018, 2020; Zhang et al. 2020; Necula et al. 2020), thus the biological interpretability of LASSO-RFE model has been confirmed.

In order to make a comprehensive comparison with the LASSO-RFE model proposed in this paper, mutual information method (MI) (Maes and Collignon 1997), L1-SVM (Li et al. 2015; Chen et al. 2021) and decision tree-based (TBS) feature selection (Chen et al. 2021; Topouzelis and Psyllos 2012; Duan et al. 2015; Yang et al. 2021) algorithms are selected as the control group in the set of Filter, Wrapper, and Embedded three major classes of feature selection algorithms, respectively, and the original data are also similarly reduced to 3 dimensions and then classified for comparison. The relevant metrics are shown in Fig. 4 and Table 5.

It is easy to see that for the four datasets involved in this experiment, RFE is the best performing feature selection algorithm among the four algorithms described and the most suitable for LASSO, both in terms of intuitive accuracy and in terms of F1 on a comprehensive consideration. In other words, RFE generally outperforms the other three algorithms in terms of combination with the LASSO algorithm.

4 Conclusion

In this paper, we propose a LASSO-RFE feature selection model based on the RFE idea and LASSO algorithm complementing each other for the nature of cancer genomic data such as high redundancy, high dimensionality, and small samples. It also consider the removal of redundant and irrelevant features and enhances the interpretability of effective gene features in a biological sense. The experimental results show that LASSO-RFE effectively reduces tens of thousands of features in the original data to three dimensions and also provides better performance for the classification model than mutual information, L1-SVM, and tree-based selection method. Meanwhile, the biological interpretability confirmed by published articles has further illustrated the accuracy of the LASSO-RFE model in feature selection. The typical applications of biometrics are to identify whether the tumor is abnormal or not, find which genes are biomarkers, diagnosis and treatment of diseases, and medicinal R&D. In this paper, only a limited number of data cases have been validated, and the application of LASSO-RFE with more recent data remains to be further investigated.

Author contributions CA contributed to conceptualization, data curation, writing—original draft, and writing—review and editing.

Funding None.

Data Availability All data are available from the corresponding author.

Declarations

Conflict of interest The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Breiman L (1995) Better subset regression using the nonnegative garrote. Technometrics 4(37):373–384

Chapman KB, Prendes MJ, Sternberg H et al (2012) COL10A1 expression is elevated in diverse solid tumor types and is associated with tumor vasculature. Future Oncol 8(8):1031–1040

Chen J, Zou Q, Li J (2021) DeepM6ASeq-EL: prediction of human N6-methyladenosine (m6A) sites with LSTM and ensemble learning. Front Comput Sci. https://doi.org/10.1007/s11704-020-0180-0

Chen K, Liu Y, Wang Z, et al (2019) Expression of COL10A1 in patients with pancreatic cancer and its prognostic value. Acad J Chin PLA Med School

Duan L, Ge H, Ma W et al (2015) EEG feature selection method based on decision tree. Bio-Med Mater Eng 26(s1):S1019–S1025

Golub TR, Slonim DK, Tamayo P et al (1999) Molecular classification of cancer: class discovery and class prediction by gene monitoring. Science 286(5439):531–537

Guyon I, Weston J, Barnhill S et al (2002) Gene selection for cancer classification using support vector machines. Mach Learn 46(1–3):389–422

Guyon I, Nikravesh M, Gunn S, et al (2006) [Studies in fuzziness and soft computing] feature extraction Volume 207|| Combining SVMs with various feature selection strategies, 315-324. https://doi.org/10.1007/978-3-540-35488-8

Huang H, Li T, Ye G et al (2018) High expression of COL10A1 is associated with poor prognosis in colorectal cancer. Onco Targets Ther 11:1571–1581

Li J, Qin Y, Yi D et al (2015) Feature selection for support vector machine in the study of financial early warning system. Qual Reliab Eng 30(6):867–877

Li Y, Wang X, Shi L et al (2020) Predictions for high COL1A1 and COL10A1 expression resulting in a poor prognosis in esophageal squamous cell carcinoma by bioinformatics analyses. Translat Cancer Res 9(1):85–94

Li T, Huang H, Shi G, et al (2018) TGF-β1-SOX9 axis-inducible COL10A1 promotes invasion and metastasis in gastric cancer via epithelial-to-mesenchymal transition. Cell Death and Disease

Maes F, Collignon A (1997) Multimodality image registration by maximization of mutual information. IEEE Trans Med Imaging 16(2):187–198

Molina LC, Belanche L, Nebot N (2002) Feature selection algorithms: a survey and experimental evaluation. In: Proceedings of the 2002 IEEE international conference on data mining (ICDM 2002), 9–12 Dec 2002, Maebashi City, Japan. IEEE.

Necula L, Matei L, Dragu D et al (2020) High plasma levels of COL10A1 are associated with advanced tumor stage in gastric cancer patients. World J Gastroenterol 26(22):3024–3033

Peng Y, Wu Z, Jiang J (2010) A novel feature selection approach for biomedical data classification. J Biomed Inform 43(1):15–23

Ramaswamy S, Golub TR (2002) DNA microarrays in clinical oncology. J Clin Oncol 20(7):1932–1941

Tibshirani R (1996) Regression shrinkage and selection via the lasso. J r Stat Soc Ser B (methodol) 58:267–288

Tinker AV, Boussioutas A, Bowtell DDL (2006) The challenges of gene expression microarrays for the study of human cancer. Cancer Cell 9:333–339

Topouzelis K, Psyllos A (2012) Oil spill feature selection and classification using decision tree forest on SAR image data. Isprs J Photogramm Remote Sens 68:135–143

Yang Y, Sun F, Chen H, Tan H, Yang L, Zhang L, Huang Y (2021) Postnatal exposure to DINP was associated with greater alterations of lipidomic markers for hepatic steatosis than DEHP in postweaning mice. Sci Total Environ 758:143631. https://doi.org/10.1016/j.scitotenv.2020.143631

Zhang M, Chen H, Wang M, Bai F, Wu K (2020) Bioinformatics analysis of prognostic significance of COL10A1 in breast cancer. Biosci Rep 40(2)

Zou Q, Xing P, Wei L, Liu B (2019) Gene2vec: gene subsequence embedding for prediction of mammalian N6-methyladenosine sites from mRNA. RNA 25(2):205–218. https://doi.org/10.1261/rna.069112.118