A New method of LncRNA classification based on ensemble learning

Zongrui Dai1, *
1Westa College, Southwest University, Chongqing, China
*dzr17723980497@gmail.com

Abstract. Long noncoding RNAs (lncRNAs), which have a length longer than 200bp (base pair), participate in various critical biological processes. Moreover, they have many similar features with another kind of RNA - coding RNA, such as long length of transcript and poly-A tail. Therefore, distinguish lncRNA and coding RNA can be an important task in bioinformatics. With the advanced and outstanding ability of machine learning, the computational method provides new insight into lncRNA classification. In this study, two feature selection methods (lasso and PCA) are applied to reduce dimension. 8 differentiated features are extracted, and lasso selection indicates better performance than the PCA method. To achieve an advanced performance of lncRNA classification, one novel ensemble learning based on primary learner and secondary learner is constructed. After comparing different kinds of models, ensemble learning achieves the most outstanding performance in AUC and accuracy within the test dataset (The median of Accuracy=0.950228, AUC=0.979664), which may shed light on the classification of lncRNA.

1. Introduction

According to the previous study, lncRNA (long non-coding RNA) is one kind of ncRNA (non-coding RNA) whose base pair number is higher than 200bp [1]. Currently, based on experiments and research in biology, the functions of lncRNA become clear. It is found that lncRNA may participate in the transcriptional and posttranscriptional regulations, which may play a role in cell expression and protein modification [2]. Moreover, the overexpression or down expression of certain lncRNA may associate with some kinds of cancers, which may serve as a biomarker or tumour suppressor [3].

Therefore, lncRNA classification is an important task in genetics and bioinformatics. To achieve advanced prediction ability, machine learning and deep learning have been widely applied. For instance, programs such as CPC2 and CNCI use SVM (support vector machine) to classify the lncRNA and coding RNA achieving advanced performance. Additionally, CPC2 mentions one novel method in feature selection which includes selection based on random forest and correlation [4-5]. Moreover, using deep learning, lncAdeep constructs one lncRNA classification system, which may have robust performance even when the RNA samples are incomplete [6]. PLEK, which combined SVM and improved k-mer scheme together, can have the ability to predict lncRNA without reference genomes [7].

In this paper, our study focuses on machine learning comparison and feature reduction. Lasso and PCA are applied in feature selection to find the suitable dimensions for the dataset. According to the results, SVM achieves the highest AUC in the testing dataset while there are 8 features both in lasso and PCA. Based on these 8 features, 5 machine learning models are compared with 10-fold-validation. The result indicates that ensemble learning (mix) achieves the highest value in AUC and accuracy. Our result
indicates that ensemble learning may achieve more robust and stable performance in testing datasets than other machine learning.

2. Methodology

2.1. Lasso regression
Lasso (Least absolute shrinkage and selection operator), which is applied widely among bioinformatics, can select potential features and decrease the dimensions. Based on Lasso regression, genes associated with cancer can be filtered, which may reduce the computational space and boost the performance of models [8]. Compared with linear regression, lasso adds one L1 regularization of weights in the loss function (Formula.1). \( \lambda \) is called the regularization parameter, which regulates the value of \( w \) (weights). With the increase of \( \lambda \), some unimportant \( w \) (weights) will be ignored. Therefore, only the important weights can be reserved, and the computational space can be reduced.

\[
loss = \frac{1}{2m} \left[ \sum_{i=1}^{m} (f(x_i) - y_i)^2 + \lambda \sum_{j=1}^{k} |w_j| \right]
\]

**Formula. 1**: Loss function of Lasso. \( m \): the number of samples, \( k \): the number of weights, \( f(x) \): the predicted value of \( y \), \( y \): the real value of \( y \), \( \lambda \): L1 regularization parameter

2.2. Ensemble learning
Ensemble learning is one kind of strategy that combines several results from different models and computes one combinational result. There are two parts in the ensemble learning - primary learner and secondary learner. In this study, the primary learner is the models of random forest, gradient boosting machine, extreme learning, and logistic regression. Taking another random forest as the secondary learner, ensemble learning can learn from the results of primary learner and give out its own prediction.

2.3. Extreme learning
The extreme learning machine is one kind of Feedforward Neural Network (FNN), which does not need to update its weights [9]. The weights of hidden neurons are randomly or artificially given at first and calculate based on an inverse matrix during the learning process [10-11]. Therefore, based on the output matrix of hidden layers(H) and the real number of y (T), the matrix of weights(\( \delta \)) can be computed at once (Formula.2).

\[
H \cdot \delta = T
\]

\[
H = \begin{bmatrix}
\delta_1 G(w_1, b, x) \\
\delta_2 G(w_2, b, x) \\
\vdots \\
\delta_{\ell} G(w_{\ell}, b, x)
\end{bmatrix}
\]

**Formula. 2**: The calculation of weights in extreme learning. \( w \) and \( \delta \) are the input weight and output weight of the hidden neurons. \( b \) is the threshold of the neurons. \( G \) is the activated function of hidden layers. \( x \) and \( T \) are the input matrix and real number of y.

3. Materials

3.1. Data sources
ENCODE (the Encyclopedia Of DNA Elements) [12], which investigates the annotation and function of the human genome, contains abundant information of lncRNA and coding RNA in the human body. The datasets of 'Protein-coding transcript sequences' and 'Long non-coding RNA transcript sequences' represented in the format of 'fasta' contains the frequency of 104,760 and 48,741.
3.2. Data processing

‘fasta’ document cannot be calculated directly through machine learning. Therefore, feature extraction and processing are needed. Using Biopython, features that have biological functions can be defined. ORFs (open-reading-frames), which are the fragments between the start codon and stop codon, represent the coding ability of RNA because the area in ORF may be reacted with enzyme and have the ability to express certain kinds of protein. Therefore, five kinds of features based on ORF are calculated by Biopython such as the length of the longest ORF (orf_length), the frequency of ORF (orf_count), the position of ORF (orf_position), the number of the start codon (start_codon_number), and the number of the end codon (end_codon_number). Besides, the transcript length of long non-coding RNA (lncRNA) has to be higher than the 200bp (base pair). Moreover, the stabilization and physical state can be different between lncRNA and coding RNA, which can be indicated by the proportion of GC pair base and pI (isoelectric point). Therefore, the features such as the length of transcripts (transcript_length), the proportion of GC pair base (GC_pro), and pI are calculated by Biopython. According to the study of CPC2, which provides additional features for IncRNA, four novel features are selected in this study such as the length of peptide (peptide_length), Fickett_score (the score, which represents the possibility of coding), and the integrity of ORF (ORF_integrity) [4]. Combined these features from Biopython and CPC2, the processed dataset contains 11 features.

3.3. Feature selection

After data processing, 11 features are calculated due to bioinformatic methods. SVM (Support Vector Machine) with RBF kernel is applied as the classification model to find out the best number of features. According to the value of AUC in each model (Table 1), when the dataset contains 8 features, the classification performance reaches the best both in Lasso and PCA. While the lasso method (AUC=0.9778747) achieves a higher AUC value than the PCA method (AUC=0.9769768). Therefore, the 8 features extracted from Lasso are selected to train the following models.

Table 1 The AUC value of each model in different number of features

| Methods | 5 features | 6 features | 7 features | 8 features | 9 features | 10 features | 11 features |
|---------|------------|------------|------------|------------|------------|-------------|-------------|
| LASSO   | 0.9691114  | 0.9717510  | 0.9765969  | **0.9778747** | 0.9774939  | 0.9774939   | 0.9776245   |
| PCA     | 0.9639177  | 0.9640641  | 0.9693370  | **0.9769768** | 0.9754171  | 0.9754171   | 0.9761841   |

4. Experiment

To gain one credible result, the 10-fold-cross validation is applied to train our models. Samples are randomly selected to be the training dataset (90%, 138151) and testing dataset (10%, 15350). Using 10-fold-cross validation, the values of AUC, Accuracy, Specificity, and Sensitivity are calculated in the test dataset. Compared with the median value of each model, ensemble learning (mix) achieves the best performance in AUC and accuracy. In contrast, extreme learning (elm) and logistic regression (glm) have the best performance in specific and sensitivity (Table 2).

Table 2 The median value of performances in test datasets

| Models | Sensitivity | Specificity | Accuracy | AUC   |
|--------|-------------|-------------|----------|-------|
| rf     | 0.960572    | 0.937101    | 0.94606  | 0.976157 |
| mix    | 0.96183     | 0.941228    | 0.950228 | 0.979664 |
| glm    | 0.965386    | 0.934326    | 0.947362 | 0.978314 |
| gbm    | 0.962564    | 0.939775    | 0.948664 | 0.977354 |
| elm    | 0.956319    | 0.944826    | 0.948827 | 0.977111 |

5. Result

The performances of SVM with RBF kernel in datasets are investigated with a different number of features. The results of lasso and PCA both indicate that SVM has the best performance while the dataset contains 8 features. Moreover, according to the AUC in each iteration, lasso selection shows a higher value of AUC than the PCA method, which implies the advanced ability of lasso regression in feature selection (Table 2). Each machine learning model is trained in R-studio with 10-fold-cross validation. The median values of each iteration are computed based on the performance in the test dataset. The
boxplot of each model indicates the performance in different kinds of assessment methods (Figure.1). No outliers are defined outside the error bar, which indicates the stable and robust performance of each model. According to the median of each model, ensemble learning achieves the best performance in accuracy (0.950228) and AUC (0.979664), while extreme learning and logistic regression perform the best in specificity (0.944826) and sensitivity (0.965386). The high accuracy and AUC of ensemble learning indicate that this method has advanced ability to classify the lncRNA and coding RNA, which may provide a new method in RNA classification.

6. Conclusion
Applying a novel assemble model, it achieves outstanding accuracy and AUC in lncRNA classification (The median of Accuracy=0.950228, AUC=0.979664), because ensemble learning can unify the performance of each weak classification model and produce one powerful model. Besides, according to our research, logistic regression and extreme learning may have the biased advantage in classifying positive samples (Sensitivity=0.965386) and negative samples (Specificity=0.944826). Therefore, under the situation of sample imbalance, logistic model and extreme learning may achieve advanced performance. At the same time, ensemble learning can balance this bias and achieve better performance in accuracy and AUC value. Our research provides one optional method, which has powerful ability in lncRNA classification, may shed light on the study of discrimination between coding and long noncoding RNA.
Reference

[1] Pauli, Andrea; Rinn, John L.; Schier, Alexander F. (2011). Non-coding RNAs as regulators of embryogenesis. , 12(2), 136–149. doi:10.1038/nrg2904

[2] Kevin C. Wang; Howard Y. Chang (2011). Molecular Mechanisms of Long Noncoding RNAs. , 43(6), 0–914. doi:10.1016/j.molcel.2011.08.018

[3] Yang, Guodong; Lu, Xiaozhao; Yuan, Lijun (2014). LncRNA: A link between RNA and cancer. Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms, 1839(11), 1097–1109. doi:10.1016/j.bbagrm.2014.08.012

[4] Kang, Yu-Jian; Yang, De-Chang; Kong, Lei; Hou, Mei; Meng, Yu-Qi; Wei, Liping; Gao, Ge (2017). CPC2: a fast and accurate coding potential calculator based on sequence intrinsic features. Nucleic Acids Research, (), –. doi:10.1093/nar/gkx428

[5] Sun, L.; Luo, H.; Bu, D.; Zhao, G.; Yu, K.; Zhang, C.; Liu, Y.; Chen, R.; Zhao, Y. (2013). Utilizing sequence intrinsic composition to classify protein-coding and long non-coding transcripts. Nucleic Acids Research, 41(17), e166–e166. doi:10.1093/nar/gkt646

[6] Yang, Cheng; Yang, Longshu; Zhou, Man; Xie, Haoiling; Zhang, Chengjiu; Wang, May D; Zhu, Huaiqiu (2018). LncADeep: An ab initio lncRNA identification and functional annotation tool based on deep learning. Bioinformatics, (), –. doi:10.1093/bioinformatics/bty428

[7] Li, Aimin; Zhang, Junying; Zhou, Zhongyin (2014). PLEK: a tool for predicting long non-coding RNAs and messenger RNAs based on an improved k-mer scheme. BMC Bioinformatics, 15(1), 311–. doi:10.1186/1471-2105-15-311

[8] Pashova H, LeBlanc M, Kooperberg C. Structured detection of interactions with the directed lasso. Statistics in Biosciences. 2017 Dec;9(2):676-691. DOI: 10.1007/s12561-016-9184-6.

[9] Guang-Bin Huang; Qin-Yu Zhu; Chee-Kheong Siew (2006). Extreme learning machine: Theory and applications. , 70(1-3), 489–501. doi:10.1016/j.neucom.2005.12.126

[10] Guang-Bin Huang, ; Qin-Yu Zhu, ; Chee-Kheong Siew, (2004). [IEEE 2004 IEEE International Joint Conference on Neural Networks (IEEE Cat. No.04CH37541) - Budapest, Hungary (25-29 July 2004)] 2004 IEEE International Joint Conference on Neural Networks (IEEE Cat. No.04CH37541) - Extreme learning machine: a new learning scheme of feedforward neural networks. , 2(), 985–990. doi:10.1109/IJCNN.2004.1380068

[11] Guang-Bin Huang; Dian Hui Wang; Yuan Lan (2011). Extreme learning machines: a survey. , 2(2), 107–122. doi:10.1007/s13042-011-0019-y

[12] Davis, Carrie A; Hitz, Benjamin C; Sloan, Cricket A; Chan, Esther T; Davidson, Jean M; Gabdank, Idan; Hilton, Jason A; Jain, Kriti; Baymuradov, Ulugbek K; Narayanan, Aditi K; Onate, Kathrina C; Graham, Keenan; Miyasato, Stuart R; Dreszer, Timothy R; Stratton, J Seth; Jolanki, Otto; Tanaka, Forrest Y; Cherry, J Michael (2018). The Encyclopedia of DNA elements (ENCODE): data portal update. Nucleic Acids Research, 46(D1), D794–D801. doi:10.1093/nar/gkx1081