A Feature Selection Method for Tissue-Specific Alternative Polyadenylation Sites Data in Rice

Wenbing Zou, Moliang Chen, Shuchao Li, Pengchao Ye and Guoli Ji*

Department of Automation, Xiamen University, Xiamen, Fujian Province, 361005, China

*Corresponding author’s e-mail: glji@xmu.edu.cn

Abstract. The identification of tissue-specific alternative polyadenylation (tsAPA) sites contributes to the research on gene expression regulation and transcriptome diversity in rice. However, identifying the tsAPA sites in plants is difficult, because of the dispersion, variability, complexity of their features and the lack of related research. A hybrid feature selection algorithm called SRBT, based on the SVM-RFE and Boruta, was presented to identify the tsAPA sites in rice. In the experiment, the tsAPA sites data were adopted to reduce dimension with SRBT algorithm and then classified by the support vector machine (SVM). The results show that the proposed method can effectively extract important features and obtain a higher average prediction accuracy of 81%, compared with SVM-RFE, Boruta, GAFS, T-test and ReliefF. The SRBT works well in identifying the tsAPA sites, which offers an effective method for further analysis of the tsAPA in gene expression and transcription during the growth of rice.

1. Introduction

As high as 70% of genes found in rice have alternative polyadenylation (APA) sites, among which there are a large number of tissue-specific Alternative polyadenylation (tsAPA) sites[1]. The APA determines the termination sites of gene transcription and plays an important role in regulating gene expression[2]. And the tsAPA affects the transcription and proteome of cells, and controls the expression of genes in specific tissues or organs[2]. The study of tsAPA contributes to finding out the mechanism of gene expression during the growth of the rice, as well as improving the rice cultivars and environmental adaptability[3].

Identifying the tsAPA sites in plants is very difficult, as a result of dispersion, variability, complexity of their features and the lack of related research[4]. There are massive features of tsAPA sites in rice, which have many redundant or irrelevant features leading to a complicated model and uncontrollable performance[5].

The traditional feature selection algorithms contain filter and wrapper, which have been widely used in bioinformatics such as sequence analysis and microarray analysis[6]. The common filtering algorithms such as T-test and ReliefF are widely used in gene expression[7],[8]. Besides, Guyon proposed SVM-RFE algorithm to study the cancer gene classification[9]. Though the filtering algorithms have the advantages of high efficiency, low computational complexity and high versatility for the large-scale datasets, the most of them have a risk to determine the optimal subset size[6]. The wrapper methods, on the contrary, generally outperform filter methods in term of prediction accuracy, but the computations are more intensive[10].
In this study, a hybrid feature selection algorithm called SBRT based on SVM-RFE[9] and Boruta[11] was presented, which combined the advantages of filter and wrapper. It can not only efficiently select the optimal features, but also automatically determine the size of feature subset. In the experiment, the important features were extracted from tsAPA sites data by SBRT. And then the SVM classification was used to distinguish tissue-specific APA sites and constitutive APA (csAPA) sites, and verify the performance based on the real biological information. It is hoped that this will be helpful for further research on the regulation of tsAPA on gene expression in rice.

2. Algorithm

2.1 SVM-RFE Algorithm

The main idea of the SVM-RFE algorithm[9] is to find an optimal classification surface, so that the classification interval on both sides of the classification surface can be maximized. In the process of eliminating features, it uses the SVM as a classifier, and then formulates specific criteria for feature ranking. However, the SVM-RFE cannot automatically determine the size of optimal feature subset.

2.2 Boruta Algorithm

Boruta algorithm[11] is a wrapper method based on the random forest classification. All the features related to variables can be captured, regardless of whether they are strongly correlated with decision variables or not. And unimportant features can be filtered recursively in the iteration process, which can minimize the errors of the model without losing some relevant features.

2.3 SRBT Hybrid Feature Selection Algorithm

The SRBT algorithm combines the advantages of SVM-RFE and Boruta algorithms, ranks the original features by recursive elimination, and then removes some redundant features with correlation information entropy to achieve the purpose of rapid reducing the dimension of the dataset. Then the Z scores of features are calculated by the random forest classifier to select the important features and obtain the optimal subset. The algorithm consists of two steps.

In the first step, the recursive elimination is used to get the ranking list of all features and the corresponding weights. Then the correlation information entropy criteria[12] is used to eliminate redundancy and obtain a reasonable size of the subsets. Correlation Information Entropy is defined as follow:

$$H_R = -\sum_{i=1}^{N} \frac{\lambda_i}{N} \log \frac{\lambda_i}{N}$$

In the formula, $\lambda_i$ represents the eigenvalue of the correlation coefficient matrix of the selected feature subsets, and $N$ is the number of features in the subsets. The bigger the $H_R$, the bigger the correlation information entropy and the smaller the correlation of the selected features. The correlation information entropy will be 0 when all features are linearly correlated, and will be 1 when all features are independent of each other.

In order to effectively remove the redundancy of the candidate subsets, the selected subsets should have the minimum redundancy. Namely, the candidate subsets have the maximum correlation information entropy.

$$\max H_R(S \cup g_i), \forall g_i \in Gene - S$$

In the second stage, we calculate the shadow features of all features in the candidate subsets and mix them together, then use random forest classifier to calculate Z scores of features in order to select important features from them. The main program of SRBT algorithm is as follow:

Algorithm SRBT:

Inputs:
Standardized dataset
\[ X = [x_1, x_2, \cdots, x_n] \]
Label set
\[ Y = [y_1, y_2, \cdots, y_n] \]
Output:
Feature subset
\[ S = [ ] \]
Step 1: Get the ranking list and corresponding weights of all features
for \( i = 0 \) to \( i = n \)
Training SVM classifier
\[ \alpha = SVM - train(X, Y) \]
Calculate weight of \( x_i \) in \( S \)
\[ w = \sum_k a_k y_k x_k, c_i = (w_i)^2 \]
Find the features of the minimum standard value and Rank
\[ f = \arg \min (c), S = [s(f), r] \]
End for
Eliminate the redundancy of \( S \) by the correlation information Entropy
for \( i = 0 \) to \( i = n' \)
Calculate \( H_R \)
\[ H_R(S \cup x_i), (x_i \in X^c - S) \]
Add the feature with the largest weight to \( S \)
\[ \text{If } (H_R(S \cup x_i) - H_R(S)) > 0 \text{, Then } S = S \cup x_i \]
End for
Step 2: Select important features from \( S \)
Calculate shadow features and blend them into \( S \)
\[ S = [(x_k, x'_k), (x_{k+1}, x'_{k+1}), \ldots] \]
Calculate Z-score by Random Forest classifier
Calculate \( Z_i \)
Find the maximum Z-score among shadow features
\[ \text{Find } MZSA = Z_{\text{max}} \]
Assign a hit to every feature that scored better than MZSA
\[ \text{If } Z_i > MZSA, h_i ++ \]
Mark the features that are less important than MZSA as "unimportant" and delete
Return \( S \)
End

3. Experiment process
The dataset used in this paper is the tissue expression data of 49196 APA sites distributed among 28,032 genes obtained from 14 tissue samples of rice (dry-seeds, embryo, endosperm, imbibed-seed, shoot, leaf-20days, leaf-60days, stem-60days, root-5days, root-60days, husk, anther, mature-pollen, pistil) after extracting APA sites and standardizing data[1][13], which has 2083 tsAPA sites and 2187 constitutive APA (csAPA) sites. There are total 226 features of the feature space in the dataset, which all belong to the RNA secondary structure, Z-curve, k-grams, nucleosome occupancy, APA context, position weight matrix (PWM), first-order heterogeneous Markov model (FHMM). The specific composition of the feature space is shown in the Table 1.
Table 1. Composition of feature Space.

| Feature Group          | Size |
|------------------------|------|
| APA context            | 7    |
| FHMM                   | 4    |
| Z-curve                | 5    |
| Nucleosome occupancy   | 4    |
| PWM                    | 4    |
| Secondary structure    | 4    |
| Conservation score     | 2    |
| k-grams                | 196  |

We randomly assigned the rice tissue data to the training and test datasets in a ratio of approximately 3:1, and performed 20 times experiments in each random process to avoid the bias of the results. In our experiments, the proposed SRBT hybrid feature selection algorithm was used to select the optimal subset, and the SVM algorithm was used for classification. In order to verify the performance of the SRBT algorithm for tsAPA sites data, we use the classical algorithms T-test[7], ReliefF[8], SVM-RFE[9], Boruta[11], GAFS[14], which are widely used to deal with biological data, to compare the performance under the same experimental conditions.

In the experiment, we use C-SVM and radial basis function in LIBSVM package. And ten-fold cross-validation was adopted to learn the optimal parameters, then the test set is evaluated with the optimal parameters, and other parameters are set to default values. The number of correctly identifying tsAPA sites was considered as a true positive ($T_p$), otherwise a false positive ($F_p$) and the number of correctly identifying csAPA sites was considered as a true negative ($T_n$), otherwise a false negative ($F_n$). The accuracy ($\frac{T_p + T_n}{T_p + T_n + F_p + F_n}$) was calculated to evaluate the performance on the classification.

4. Analysis of experimental results

From Table 2, we note that compared with the other five algorithms, the SRBT hybrid feature selection algorithm has the smallest dimension of subset after dealing with the tsAPA sites data, and the feature space is reduced from 226 to 10. At the same time, it has the highest average prediction accuracy, reaching 81%.

Table 2. Performance comparison of the feature selection methods.

| Method   | Size | Accuracy       |
|----------|------|----------------|
| T-test   | 89   | 0.6573±0.0510  |
| ReliefF  | 72   | 0.6724±0.0631  |
| GAFS     | 62   | 0.7126±0.0379  |
| SVM-RFE  | 226  | 0.6959±0.0415  |
| Boruta   | 53   | 0.7383±0.0230  |
| SRBT     | 10   | 0.8110±0.0352  |

Table 3 shows that the APA context has the most features (Adis_front, Adis_stop_codn, Adis_S, Adis_E) in the optimal subset obtained by SRBT algorithm, which belong to the genetic information features of APA sites. The sequence and distribution of the np_max (nucleosome occupancy) plays an important role in the formation of complex DNA structures. The SmaxN and SmeanR are single-stranded regions formed by the complementation of four bases in the RNA sequence. And the other selected features are the important sequence composition of APA sites. We can study the inner mechanism between tsAPA and gene expression in rice by in-depth analysis of these features.

Table 3. Final selected features.

| Feature         | Feature group | Description                |
|-----------------|---------------|----------------------------|
| Adis_front      | APA context   | The distance to 5' APA site |
| Adis_stop_codon | APA context   | The distance to the stop codon |
5. Conclusion

This study presents a hybrid feature selection algorithm SRBT based on SVM-RFE and Boruta for identifying tsAPA sites in rice. Compared with SVM-RFE, Boruta, GAFS, T-test and ReliefF, the results show that the SRBT works better in reducing dimension and identifying the tsAPA in rice, and achieves the highest average prediction accuracy of 81%. The features selected by the SRBT, can affect the real transcription process of the genes and have better performance on classification, which contribute to the further research on the regulation of APA sites on gene expression in rice.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Nos.61573296, 61802323, 61871463 and 61803320).

References

[1] H. Fu et al. (2016) Genome-wide dynamics of alternative polyadenylation in rice. Genome Res, vol. 26, no. 12, pp. 1753–1760.
[2] B. Tian and J. H. Graber. (2012) Signals for pre-mRNA cleavage and polyadenylation. Wiley Interdiscip. Rev. RNA, vol. 3, no. 3, pp. 385–396.
[3] Y. Shen et al. (2008) Genome level analysis of rice mRNA 3′-end processing signals and alternative polyadenylation. Nucleic Acids Res, vol. 36, no. 9, pp. 3150–3161.
[4] G. Ji, J. Guan, Y. Zeng, Q. Q. Li, and X. Wu. (2015) Genome-wide identification and predictive modeling of polyadenylation sites in eukaryotes. Brief. Bioinform, vol. 16, no. 2, pp. 304–313.
[5] S. Dudoit, J. Fridlyand, and T. P. Speed. (2002) Comparison of discrimination methods for the classification of tumors using gene expression data. J. Am. Stat. Assoc, vol. 97, no. 457, pp. 77–86.
[6] Y. Saeys, I. Inza, and P. Larrañaga. (2007) A review of feature selection techniques in bioinformatics. Bioinformatics, vol. 23, no. 19, pp. 2507–2517.
[7] M. R. Ingrid Hedenfalk , David Duggan et al. (2001) Gene-expression profiles in hereditary breast cancer. Adv. Anat. Pathol, vol. 344, no. 8, pp. 1–6.
[8] I. Kononenko. (2002) Estimating attributes: Analysis and extensions of RELIEF. Machine Learning: ECML-94, 1994, pp. 171–182.
[9] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik. (2002) Gene Selection for Cancer Classification using Support Vector Machines. Mach. Learn, vol. 46, no. 1, pp. 389–422.
[10] Z. Zhu, Y. S. Ong, and M. Dash. (2007) Wrapper-filter feature selection algorithm using a memetic framework. IEEE Trans. Syst. Man, Cybern. Part B Cybern., vol. 37, no. 1, pp. 70–76.

[11] W. R. Rudnicki, M. Kierczak, J. Koronacki, and J. Komorowski. (2006) A Statistical Method for Determining Importance of Variables in an Information System. Rough Sets Curr. Trends Comput., vol. 4259, pp. 557–566.

[12] Q. Wang, Y. Shen, Y. Zhang, and J. Q. Zhang. (2003) A quantitative method for evaluating the performances of hyperspectral image fusion. IEEE Trans. Instrum. Meas., vol. 52, no. 4, pp. 1041–1047.

[13] G. Ji et al. (2018) TSAPA: Identification of tissue-specific alternative polyadenylation sites in plants. Bioinformatics, vol. 34, no. 12, pp. 2123–2125.

[14] H. Chouaib, O. Terrades, S. Tabbone, F. Cloppet, and N. Vincent. (2008) Feature selection combining genetic algorithm and adaboost classifiers. Proceedings - International Conference on Pattern Recognition, pp. 1–4.