Application of Machine Learning in Analysis of Transcriptomic Data Derived from Next Generation Sequencing

Meng-Hsiun Tsai, Shien-Chung Huang, Hsin-Hung Yeh and An-Yuan Chu

November 18, 2019

Application of Machine Learning in Analysis of Transcriptomic Data Derived from Next Generation Sequencing

Meng-Hsiun Tsai
Natio`sity, Taiwan
145 Xingda Rd., South Dist.,
Taichung City 402, Taiwan (R.O.C.)
+886-4-22840864 ext. 666
mht@nchu.edu.tw

*Shien-Chung Huang
National Chung Hsing University, Taiwan
145 Xingda Rd., South Dist.,
Taichung City 402, Taiwan (R.O.C.)
+886-4-22840230 ext. 310
hchwang@dragon.nchu.edu.tw

Hsin-Hung Yeh
ABRC, Academia Sinica, Taiwan
No. 128, Sec. 2, Academia Road, Nankang, Taipei 115, Taiwan (R.O.C.)
+886-2-27872116
hyeh@sinica.edu.tw

An-Yuan Chu
National Chung Hsing University, Taiwan
145 Xingda Rd., South Dist.,
Taichung City 402, Taiwan (R.O.C.)
+886-4-22840230 ext. 310
breakingwithme@gmail.com

ABSTRACT

Tobacco Mosaic Virus, the most studied plant virus, could infect over 100 species of plants and over 550 species of flowering plants, causing enormous loss of economy at home and abroad. Microarray, an important analytic tool of Genomics and Genetics, enables researchers to analyze massive gene expression simultaneously. To find out the genes related to replication of the Tobacco Mosaic Virus, the material of this research is gene expression of the cell of Arbidopsis infected by Tobacco Mosaic Virus, which recorded in 5 time points (30 min, 4hr, 6hr, 18hr and 24hr) and made by Next Generation Sequencing. The research analyzes the time-series raw data and adapts the Fast Correlation-Based Filter (FCBF) and the Wrapper algorithms for gene selection. The selected genes are validated by the C4.5 algorithm and Multi-Layer Perceptron. Results show that genes selected by Wrapper algorithm with average accuracy 75%, average true positive rate (classified accuracy of control group) 77.5%, true negative rate (classified accuracy of experiment group) 72.5%, average F-measure 74.85% and average AUC 07965, perform better overall than genes selected by other algorithms.

Keywords

Tobacco Mosaic Virus ; Next Generation Sequencing ; Arbidopsis ; Machine Learning

1. INTRODUCTION

Tobacco Mosaic Virus (TMV) is the most studied positive-sense single-stranded RNA (+ssRNA) plant virus. In 1935, Wendell Meredith Stanley crystallized TMV, which normally includes big changes in temperature, pressure, etc. to inactivate matters. Surprisingly, TMV could remain activation after being crystallized [1]. Due to the large host range of TMV, including over 100 species of plants and over 550 species of flowering plants, researchers can inoculate it with other plants with obvious characterizations, like Arbidopsis, to find out how the virus affects plants [2, 3].

Microarray, a collection of microscopic DNA spots attached to a solid surface, applied by scientists to measure the expression of large numbers of genes simultaneously [4]. The high demand for low-cost sequencing has driven the development of high-throughput sequencing, which also goes by the term next generation sequencing (NGS). Thousands or millions of sequences are concurrently produced in a single next-generation sequencing process. With the commercialization of various affordable desktop sequencers, NGS has become within the reach of traditional wet-lab biologists. In recent years, genome-wide scale computational analysis is increasingly being used as a backbone to foster novel discovery in biomedical research. However, as the quantities of sequence data increase exponentially, the analysis bottle-neck is yet to be solved [5, 6, 7, 8].

The research aims to apply machine learning algorithms to analyze massive genes expression datasets, construct an easier and more efficient analysis model, and find out important biomarkers.

Section 2 reviews previous works on feature selection and supervised learning algorithms. Section 3 describes the materials and methods used in the research. Section 4 presents the experiment process and results. Finally, Section 5 wraps up with the main conclusions.

2. Literature Review

2.1 Feature Selection

In analysis of gene expression data, the researchers have to face the curse of dimensionality, which means the data structure consists of massive attributes and few samples. Feature selection, which can identify and delete irrelevant and redundant attributes, is applied to overcome the problem [9, 10].

(1) Filter methods

The most researched methods with applying statistics methods to calculate metrics of evaluating one attribute or attribute subset. The advantages of filter methods are the speed of calculation and low cost of resource by discretizing and simplifying data before calculation. The disadvantages are the lack of consideration of the relationship between attributes and no connection with later learning algorithms, which lead to inferior results [11].

Fast Correlation-Based Filter (FCBF), proposed by Yu and Liu, is modification of original linear algorithms. Symmetrical Uncertainty (SU) [12] is calculated for non-linear real data (1).

$$SU(X,Y) = 2\left[\frac{IG(Y)}{H(X)+H(Y)}\right]$$

$H(X)$, $H(Y)$ is entropy of X and Y, $IG(Y)$is gain of Information Gain. If $SU = 1$, X is fully related to Y, vice versa.

(2) Wrapper methods

Wrapper first tests every possible subset by applying a learning algorithm to evaluate, and after calculating the error rate of every subset, ranks the subset and picks out the best. The advantages are the accuracy and interact with learning algorithm; however, lots of iterations cost length time of calculation and plenty of resources.

2.2 Supervised learning

Each sample of the dataset have corresponding target variables, the algorithms learn from historical data to find out the best pattern and construct model. Three common types of supervised learning are as follows.

(1) Regression is to predict numerical target variables.

(2) Classification is to predict categorical target variables.

(3) Anomaly detection is to find out the abnormal data point.

i. Decision tree, the most common supervised learning algorithm, is derived step by step until every sample is classified. The best scenario is the samples, which classified to a leaf node, from same class of raw data.

ii. Multi-Layer perceptron, consisted of many perceptron, includes input layer, hidden layer and output layer [13].

3. Materials and Methods

3.1 Data resource

The dataset, Tobacco Mosaic Virus (TMV) infected-transcriptomic data derived from Next Generation Sequencing (NGS), is from Academia Sinica of Taiwan. The dataset consists of 41,671 gene expressions for each time point, recorded as Transcript per Million (TPM).

TMV-Rep* set as the control group is a chemically treated virus and cannot replicate, TMV-U1set as the experimental group is the wild virus. The gene expression was recorded with 0.5 hour as mock. (gene expression only affected by experimental operation, such as transfer buffer and enzyme impact). 4 hours and 6 hours after infected record as the cumulative rising of the quantity of virus; 18 hours and 24 hours after infected the stability of the quantity of virus.

3.2 Data preprocessing

To fulfill the datatype limitation of FCBF and C4.5, the numerical attributes should be discretized to categorical attributes. Hence, in consideration of consistency, the research normalized the values of each time point to [0, 1] for each gene. Based on the four changes between five time points, regroup 4 groups with an interval of 0.25 for each group.

3.3 Feature Selection – Gene Selection

Due to the raw dataset is consist of pure numerical attributes and categorical target variable, the research not only discretized and data transformed to fulfill some feature selection algorithms, but also remain the probability to select features with numerical attributes of raw dataset.

(1) To achieve the best time efficient and for discretized data, the research chooses the FCBF algorithm of filter methods to select genes. At first, discretize the attributes. Second, calculate the SU of each attribute to target variables. The SU is between 0 to 1, 0 means that the attribute has no relation with target variables, vice versa. Third, calculate the SU between each attribute. In this step, the selected attributes from the second step will be set as the target variable in turns to pick out redundant and repeated genes. The inputs of algorithms should be discretized attributes and corresponding target variables and the outputs genes with the biggest SU (threshold depends on the operator).

(2) To achieve the best accuracy and for raw numerical data, the research chooses the Wrapper method to select genes. Attributes should remain numerical after preprocessing, and then the searching algorithm and learning algorithm should be chosen. For the searching algorithm, the research chooses the Best-First algorithm, which adopts the structure of priority ranking with choosing the best part for each time until test all the possible subsets. As for the learning algorithm, MLP is chosen.

3.4 Learning algorithm – Verification of selected genes

To verify the genes selected by previous feature selection, the research use learning algorithms C4.5 Decision Tree and Multi-Layer Perceptron for either categorical attributes or numerical attributes to estimate every selected genes subset and construct classification model.

As for the C4.5 Decision Tree, SU of the FCBF applied, in consideration of non-linear problems to handle real data more appropriately. In Multi-Layer Perceptron, the research adjusts the input and output layer based on the number of genes from each feature selection algorithm. The nodes of the input layer are the raw data of each gene, the nodes of the output layer two nodes from the control and experimental group. The nodes of the hidden layer will be the sum of nodes of the input layer and output layer divided by 2, the learning rate is 0.3 and the stop criteria is set to 95% accuracy.

The verified results will be shown as the comparison of Sensitivity, Specificity, Accuracy, F-measure, and AUC for both algorithms.

4. Experiment and Results

4.1 Experiment Design and Process

The experiment process, showed as follows (Fig 1.), includes four parts, which are data preprocessing, gene selection, verification of selected genes, and gene function exploration. R (v3.5.1) and WEKA (v3.8.1) are applied in the research with Windows 10.

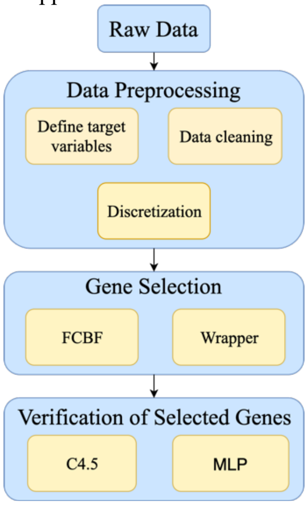

Fig 1. Experiment process

Table 1. Structure of dataset

	0.5 hr (mock)	4 hr	6 hr	18 hr	24 hr
TMV-Rep*	KT01-1/2	KT02-1/2	KT03-1/2	KT04-1/2	KT05-1/2
	KT11-1/2	KT12-1/2	KT13-1/2	KT14-1/2	KT15-1/2
TMV-U1	KT06-1/2	KT07-1/2	KT08-1/2	KT09-1/2	KT10-1/2
	KT16-1/2	KT17-1/2	KT18-1/2	KT19-1/2	KT20-1/2

4.2 Experiment results

4.2.1 Data Preprocessing

The results of data preprocessing will be shown in four parts as follows.

(1) Define target variables.

Based on the structure of raw dataset and referring papers, the research defines two target variables groups, with and without considering chronical issue. The research reverses attributes and target variables, which set genes as samples and original target variables as attributes of genes. The details are explained in data transformation part.

(2) Data cleaning.

Delete genes with no value at multiple samples. After observation and statistics, 4,183 genes with no value at every sample and 12,525 genes with no value over half samples (20 samples) are removed from the raw dataset, and ultimately, 24,963 genes have remained.

(3) Discretization.

The data is discretized after data cleaning to fulfill the calculation of the SU value of FCBF and the limitation of the C4.5 decision tree.

4.2.2 Gene Selection

The research applies FCBF and Wrapper to select genes based on different target variables' definitions.

(1) The genes selection results of FCBF.

i. Without considering chronical issue, divide to control and experimental group.

After calculating all SU of each gene to target variables, the maximum SU is 0.3461, belongs to Transcript ID AT5G53550.1, and the minimum is 0 with 1,407 genes in total. For only considering control and experiment group, the relationship between genes and target variables is very low (the maximum SU < 0.35). Even though, the research picks out 14 genes (Table 2.) with SU over 0.2, calculating the relationship between each gene as same as above and picking out the minimum average SU. 4 genes with average SU less than 0.25 are picked out including AT2G23880.1, AT2G10770.1, AT5G15340.1 and AT1G73210.2 (Table 3.).

ii. With considering chronical issue, remain one dataset, 10 classes.

After calculating all SU of each gene to target variables, the maximum SU is 0.774003 with 5 genes in total, and the minimum SU is 0.0968, which belongs to Transcript ID AT5G02950.1. The relationship between genes and target variables is higher in consideration of the chronical issue, compared to without considering the chronical issue (the maximum SU > 0.75); hence, the research picks out 66 genes with SU over 0.73. (Table 4.) However, there are multiple genes are redundant. After calculation, 9 genes with average SU less than 0.8 are picked out, include AT4G34120.1, AT3G04400.1, AT3G07110.1, AT3G53870.1, AT4G17390.1, AT4G02940.1, AT5G01720.1, AT3G46290.1 and AT5G09900.1 (Table 5.). Within 9 genes above, there are 4 genes and 2 genes with average SU 0.7204 and 0.7794, which means they are redundant to each other. The research chooses the one with a larger SU to target variables among them.)

iii. With considering chronical issue, divide to two datasets, 5 classes for each.

Because of dividing into two datasets, the research can only calculate the difference of SU to pick out genes, which are expressed differently in two situations. The maximum SU difference is 0.6429, which belongs to Transcript ID AT4G31660.1, and the minimum SU difference is 0, with a total of 383 genes. Compare to other definitions, the SU are higher and the redundant genes are lesser. As a result, the research first chooses 7 genes with SU difference are higher than 0.6 (Table 6.), and after calculating the relationship between each gene, 4 genes with average SU less than 0.26, including AT5G20870.1, AT3G41768.1, AT3G22121.1 and AT1G09190.1 (Table 7.) are picked out.

(2) The genes selection results of Wrapper.

i. Without considering chronical issue, divide to control and experiment group.

The research picked out 5 genes include AT1G12090.1, AT1G27060.1, AT3G07470.1, AT3G18480.1, AT3G27110.1.

ii. With considering chronical issue, remain one dataset, 10 classes.

The research picked out 3 genes include AT1G55680.1, AT4G19880.1, AT4G24000.1.

iii. With considering chronical issue, divide to two datasets, 5 classes for each.

The research picked out one gene for the control group dataset and experimental group dataset separately, which is the control group dataset, AT1G76150.1, and the experimental group dataset, AT1G45000.1.

From the above, the number of genes as the best subset will change with different definitions. Although Wrapper can find out genes more precisely than FCBF, calculation time and cost are heavier.

4.2.3 Verification of Selected Genes

Totals of 29 genes will be verified by C4.5 and MLP group by group. The Sensitivity is the ratio that correctly classified as the experimental group, and the Specificity is the ratio that correctly classified as the control group.

(1) Genes from FCBF.

In C4.5, the accuracy is 65%, the sensitivity is 75%, the specificity is 55%, the F-measure is 63.46% and the AUC is 0.646. In MLP, the accuracy is 62.5%, the sensitivity is 65%, the specificity is 60%, the F-measure is 62.4% and the AUC is 0.628.

(2) Genes from Wrapper.

In C4.5, the accuracy is 70%, the sensitivity is 70%, the specificity is 70%, the F-measure is 70% and the AUC is 0.746. In MLP, the accuracy is 80%, the sensitivity is 85%, the specificity is 75%, the F-measure is 79.7% and the AUC is 0.847.

As a result, genes from FCBF perform better in C4.5, whereas genes from Wrapper perform better in MLP but perform better than genes from FCBF in C4.5.

Table 2. FCBF without considering chronical issue, ranks SU between genes and target variables.

Transcript ID (Gene)	SU
AT5G56550.1	0.3461
AT2G04040.1	0.2336
AT1G73210.2	0.2303
AT5G60700.1	0.2303
AT1G27020.1	0.2303
AT3G15450.1	0.2162
AT5G10960.1	0.2134
AT5G26920.1	0.2098
AT2G23880.1	0.2081
AT5G06090.1	0.2081
AT5G15340.1	0.2068
AT2G10770.1	0.2040
AT2G17740.1	0.2021
AT5G01640.1	0.2002

Table 3. FCBF without considering chronical issue, ranks average SU between genes.

Transcript ID (Gene)	Average SU
sAT2G23880.1	0.1858
AT2G10770.1	0.2002
AT5G15340.1	0.2350
AT1G73210.2	0.2400
AT3G15450.1	0.2695
AT5G01640.1	0.2964
AT5G06090.1	0.3088
AT1G27020.1	0.3397
AT5G60700.1	0.3633
AT5G10960.1	0.4093
AT1G56550.1	0.4334
AT5G26920.1	0.4411
AT2G17740.1	0.4556
AT2G04040.1	0.4658

Table 4. FCBF with considering chronical issue and remain one dataset, ranks SU between genes and target variables.

Transcript ID (Gene)	SU
AT3G04400.1 AT3G07110.1 AT3G53870.1 AT4G02940.1 AT4G17390.1 AT3G09390.1	0.745
AT1G01800.1 AT1G02220.1 AT1G04480.1 AT1G08480.1 AT1G10370.1 AT1G10450.1 AT1G14320.1 AT1G15270.1 AT1G20225.1 AT1G20450.1 AT1G27970.1 AT1G30230.1 AT1G32210.1 AT1G53850.1 AT1G56450.1 AT1G62740.1 ⋮ ⋮	0.733

Table 5. FCBF with considering chronical issue and remain one dataset, ranks average SU between genes.

Transcript ID (Gene)	SU
AT4G34120.1	0.6739
AT3G04400.1 AT3G07110.1 AT3G53870.1 AT4G17390.1	0.7204
AT4G02940.1	0.7297
AT5G01720.1	0.7497
AT3G46290.1 AT5G09900.1	0.7204

Table 6. FCBF with considering chronical issue and divide to two datasets, ranks SU difference between genes and target variables.

Transcript ID (Gene)	SU
AT4G31660.1	0.6429
AT4G14860.1	0.6328
AT3G22121.1	0.6260
AT3G41768.1	0.6239
AT1G09190.1	0.6206
AT2G18720.1	0.6091
AT5G20870.1	0.6007

Table 7. FCBF with considering chronical issue and divide to two datasets, ranks average SU between genes.

Transcript ID (Gene)	Average SU
AT5G20870.1	0.2238
AT3G41768.1	0.2367
AT3G22121.1	0.2504
AT1G09190.1	0.2586
AT2G18720.1	0.2743
AT1G14860.1	0.2757
AT1G31660.1	0.2977

Table 8. Verification of Selected Genes

		FCBF	Wrapper
Without considering chronical issue, divide to control and experiment group.		AT2G23880.1 AT2G10770.1 AT5G15340.1 AT1G73210.2	AT1G12090.1 AT1G27060.1 AT3G07470.1 AT3G18480.1 AT3G27110.1
With considering chronical issue	One dataset	AT4G34120.1 AT3G04400.1 AT4G02940.1 AT5G01720.1 AT3G46290.1	AT1G55680.1 AT4G19880.1 AT4G24000.1
	Two datasets	AT5G20870.1 AT3G41768.1 AT3G22121.1 AT1G09190.1	AT1G76150.1 AT1G45000.1

5. Conclusion

The research aims to analyze Transcriptomic data derived from Next Generation Sequencing, find out the genes relate to replicate mechanism of virus and construct analysis model. The research chooses FCBF and the Wrapper algorithms for gene selection, and verify the results with C4.5 and MLP. Genes selected from the Wrapper algorithm perform better overall than the ones from FCBF algorithms. Previous papers related to applying machine learning to analyze gene expression data almost views genes as features of diseases, genders, etc. For further study may attempt to set genes as samples and transforms new attributes from the original target variables to overcome the disadvantages of the FCBF and the Wrapper algorithms.

6. REFERENCES

[1] Zaitlin, M. (1998). The discovery of the causal agent of the tobacco mosaic disease. In *Discoveries In Plant Biology: (Volume I)* (pp. 105-110).

[2] Shors, T. (2016). Understanding viruses. Jones & Bartlett Publishers.

[3] Sanfaçon, H. (2017). Grand challenge in plant virology: understanding the impact of plant viruses in model plants, in agricultural crops, and in complex ecosystems. *Frontiers in microbiology*, *8*, 860.

[4] Tarca, A. L., Romero, R., & Draghici, S. (2006). Analysis of microarray experiments of gene expression profiling. *American journal of obstetrics and gynecology*, *195*(2), 373-388.

[5] Mardis, E. R. (2008). Next-generation DNA sequencing methods. *Annu. Rev. Genomics Hum. Genet.*, *9*, 387-402.

[6] Kukurba, K. R., & Montgomery, S. B. (2015). RNA sequencing and analysis. *Cold Spring Harbor Protocols*, *2015*(11), pdb-top084970.

[7] Miller, M. B., & Tang, Y. W. (2009). Basic concepts of microarrays and potential applications in clinical microbiology. *Clinical microbiology reviews*, *22*(4), 611-633.

[8] Bumgarner, R. (2013). Overview of DNA microarrays: types, applications, and their future. *Current protocols in molecular biology*, *101*(1), 22-1.

[9] Guyon, I., Gunn, S., Nikravesh, M., & Zadeh, L. A. (Eds.). (2008). *Feature extraction: foundations and applications* (Vol. 207). Springer.

[10] Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of machine learning research*, *3*(Mar), 1157-1182.

[11] Hall, M. A. (1999). Correlation-based feature selection for machine learning.

[12] Yu, L., & Liu, H. (2003). Feature selection for high-dimensional data: A fast correlation-based filter solution. In *Proceedings of the 20th international conference on machine learning (ICML-03)*(pp. 856-863).

[13] Gardner, M. W., & Dorling, S. R. (1998). Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences. *Atmospheric environment*, *32*(14-15), 2627-2636.