Int. J. Agricult. Stat. Sci. Vol. 17, Supplement 1, pp. 2419-2426, 2021
DocID: https://connectjournals.com/03899.2021.17.2419
www.connectjournals.com/ijass
ISSN : 0973-1903, e-ISSN : 0976-3392

ORIGINAL ARTICLE

ALGORITHM FOR SELECTION OF INFORMATIVE GENES USING GENE EXPRESSION DATA

Nitesh Kumar Sharma*, Dwijesh Chandra Mishra, Mohammad Samir Farooqi, Neeraj Budhlakoti, Krishna Kumar Chaturvedi, Samrendra Das, Anil Kumar and Anil Rai

Centre for Agricultural Bioinformatics, ICAR-Indian Agricultural Statistics Research Institute, New Delhi - 110 012, India.
E-mail: sharmanitesh.iasri@gmail.com

Abstract: Informative gene selection from high dimensional gene expression data has appeared as an important area of research in agri-genomics. Different gene selection techniques have been developed in recent times based on relevancy and redundancy of genes with class and among the genes. Most popular techniques for informative gene selection are Maximum Relevancy and Minimum Redundancy (MRMR) and Support Vector Machine Recursive Feature Elimination (SVM-RFE). However, these methodology have some drawbacks. One of the major drawback is that **it** ignores the spurious relations between genes and trait under study. In this study, a methodology for informative gene selection has been developed, which takes care of this spurious relation by implementing the bootstrap technique along with SVM-RFE and MRMR. The performance of these gene selection techniques has been analysed through classification accuracy of the SVM model with linear kernel developed using selected informative genes as predictors. A comparative evaluation of the developed method was done against three well known existing techniques for gene selection *viz.* Boot-MRMR, SVM-RFE, MRMR. On the basis of various evaluation measures, it has been observed that the performance of the developed methodology is better as compared to above given techniques and select less number of more informative genes. Moreover, for proper implementation and dissemination of the developed methodology, a user friendly R software package named "IGST" has been developed by using state of the art technology.

Key words: Bootstrapping, Gene expression, Informative gene, MRMR, SVM-RFE.

Cite this article

Nitesh Kumar Sharma, Dwijesh Chandra Mishra, Mohammad Samir Farooqi, Neeraj Budhlakoti, Krishna Kumar Chaturvedi, Samrendra Das, Anil Kumar and Anil Rai (2021). Algorithm for selection of informative genes using gene expression data. *International Journal of Agricultural and Statistical Sciences.* DocID: https://connectjournals.com/03899.2021.17.2419

1. Introduction

Information of genes has high feasibility and necessity in enhancing agricultural production. Selecting a parsimonious set of informative genes is one of the most important tasks for the analysis of gene expression data, as it can help to determine the underlying mechanism of several traits or stress related functions. Selection of informative genes from a set of very large number of genes can be viewed as selecting pertinent features, which could better classify the observations into two categories or classes as for example diseased or non-diseased class [Diaz-Uriarte (2007), Díaz-Uriarte and De Andre (2006)]. It is tough to decide which genes are valuable without prior knowledge. Therefore, many numbers of genes are generally introduced into the dataset, including redundant, relevant and irrelevant genes. However, redundant and irrelevant genes are not advantageous for classification [Yu and Liu (2004)]. To select a subset of the genes from the original gene expression data, the gene selection procedure uses an optimization algorithm which provides a subset of informative genes having the most classification information [Pyingkodi and Thangarajan (2017), Zhu *et al.* (2010)]. There are numerous methods for informative genes selection like different test, fold chain methods, feature selection methods *etc.* Among

*Author for correspondence

Received *March 17, 2021* ***Revised*** *July 14, 2021* ***Accepted*** *August 29, 2021*

them, feature selection is best approach to be measured [Dash and Liu (1997)].

Informative gene (features) selection techniques like MRMR [Ding and Peng (2005), SVM-RFE [Guyon *et al.* (2002), SVM-RFE with MRMR [Mundra and Rajapakse (2009), Tang *et al.* (2007)] *etc.* are able to provide the informative and useful features but also have some drawbacks associated with them. The MRMR filter, when used alone, may not yield optimal accuracy because the classifier performs independently and is not involved in the selection of genes [Guyon and Elisseeff (2003)]. On the other hand, SVM weights in SVM-RFE provide a criterion to rank genes based on their relevancy, but they do not account for the redundancy among the genes. Further, a combination of SVM-RFE with MRMR was used to improve classification accuracy and to minimize redundancy among relevant genes [Liang *et al.* (20110], but this combination also does not care about the spurious associations of the genes with the target trait/conditions as well as among other genes. Such spurious association may be minimized by applying bootstrapping procedure. Das *et al.* (2017) gave a statistical method *i.e.* Bootstrap SVM-RFE based on support vector machine algorithm for selecting informative genes from high dimensional gene expression data [Das *et al.* (2017, 2018)]

. In this paper, we describe a new method developed for feature selection, which uses bootstrap samples in the process of gene selection through SVM-RFE along with MRMR filter.

2. Materials and Methods

2.1 Data Collection

The GE experimental dataset under abiotic stress (salinity) for Rice was collected from Gene Expression Omnibus database of NCBI for platforms GPL2025 (www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GPL2025), as this platform contains as much as 191 microarray experiments (series) comprising 3096 samples/subjects of *Oryza sativa* L. as compared to other platforms. The pre-processing of the gene expression datasets was done to remove noises, including missing probes and mislabelled probes [Das *et al.* (2017)]. Here, the pre-processing of data was conducted by using Bioconductor platform of R [Gentleman *et al.* (2004)]. Initially, the raw CEL files of the collected samples were processed using Robust Multichip Average (RMA) algorithm available in *affy* Bioconductor package of R [Gautier *et al.* (2004), Bolstad *et al.* (2003)]. This RMA procedure involves background correction, quantile normalization and summarization by median polish approach. Further, the $\log_2$ scale transformed expression data from RMA for the collected experimental samples were used for meta-analysis to remove the outlier samples.

2.2 Bootstrapping

Here, we propose a technique, which uses a modified bootstrap based subject sampling model for selection of informative genes through SVM-RFE with MRMR filter. In this model, the gene expression (GE) experimental samples *i.e.* subjects are taken as sampling units from the population. Each subject has GE measurement for same set of genes. Moreover, the replicated GE sample is taken as new sampling units under this sampling model, which may have different GE profiles as compared to other replicates [Goeman and Bühlmann (2007)].

Let, M denotes population size, *i.e.* total number of GE profiles for different subjects in the experiment and each subject is treated as an independent unit in the population. The relation of each subject with its class can be shown as

$$(X_1, y_1), (X_2, y_2), ..., (X_s, y_s), ..., (X_M, y_M) \qquad (1)$$

where, X_s represents the N- dimensional vector (N is total number of genes) of GE levels for sth subject and y_s is the corresponding class label (*e.g.* stress: +1 *vs.* control: –1), $s = 1, 2, \ldots, M$. Therefore, M expression levels of different subjects/samples are independently and identically distributed (iid), but expression levels of genes within the same subject may be correlated for a given condition. Let, M units be randomly selected with replacement to construct one bootstrap sample. Then, the SVM with MRMR filter algorithm is applied on this bootstrap sample to get one list of the genes along with their ranks (say one gene list). This procedure is repeated S times to get S gene lists. Here, S, *i.e.* number of bootstrap samples must be sufficiently large [Wang *et al.* (2013)]. It has been empirically established that value of S should be around 200 to ensure all desirable features of bootstrapping [Efron and Tibshirani (1994)].

2.3 MRMR

The MRMR (Maximum Relevance and Minimum Redundancy) method aims at selecting maximally relevant and minimally redundant set of genes for discriminating classes. In usual MRMR technique, genes are ranked by optimizing the combination of relevance and redundancy measures under following schemes *i.e.* F-test with correlation difference (FCD) and F-test with correlation quotient (FCQ), mutual information measures [Ding and Peng (2005), Peng *et al.* (2005)]. In this method, we applied F-test with mutual information measure.

Let, the MRMR objective function (J) for the gene selection problem be given as

$J = max(V/W)$

where, V and W denote the relevance and redundancy measures, respectively. Further, the near optimal solution of J for the GE data is obtained by Ding and Peng (2005) and is given as

$$w_i = \max_{i\in\Omega}\left\{F(i,y)\Big/\left(\frac{1}{|\Omega|}\sum_{i\neq k=1}^{|\Omega|}|I(i,k)|\right)\right\} \tag{2}$$

where, Ω is the gene space, w_i is the weight associated with *i*th gene, y is the class label of a subject, $F(i,y)$ is the F-score between the *i*th gene in y class and $I(i, k)$ is mutual information measure between *i*th and *k*th gene in GE dataset. This weight is used in the calculation of final ranking measure.

2.4 SVM

Support vector machines (SVMs) [Vapnik (1998)] are supervised learning models with associated algorithms that analyse the data for the purpose of classification and regression analysis. The component of the weight vector v of the SVM as follows.

$$v = \sum_s \alpha_s y_s x_s \tag{3}$$

where, y_s is the class label of the sample X_s and the summation is taken over all the training samples. α_s is the Lagrange multipliers involved in maximizing the margin of separation of the classes. This weight vector is used in final ranking measure for selection of informative genes.

2.5 SVM with MRMR filter

In our approach, SVM along with MRMR filter algorithm ranks the genes by a combination of the relevancy given by SVM weights and the MRMR criterion which provides a list of genes along with their ranks for each bootstrap sample. For *i*th gene, the ranking measure ρR_i is as follow.

$$R_i = \beta|v_i| + (1-\beta)w_i \tag{4}$$

The parameter $\beta \in [0, 1]$ determines the trade-off between SVM weight and MRMR weight. By applying this algorithm on each bootstrap sample, we consequently get a gene list with their rank. Thus, many gene lists are obtained and the number of such gene lists is equal to number of bootstrap samples.

2.6 Statistical Framework

Further, genes in every gene list have ranks between 1 to N. Then a function, *i.e. RankScore* for *j*th gene list, *i.e.* $R_j(R_j : \Omega \rightarrow [0,1])$ is defined to map ranks of genes to corresponding scores [Dasgupta *et al.* (2018)]. The *Rank Score* $\left(R_j^{(i)}\right)$ for *i*th gene in *j*th ($k = 1, 2, \ldots,$ S) gene list can be defined as

$$R_j^{(i)} = f(\rho_{ij}) = \frac{N-\rho_{ij}+1}{N} \tag{5}$$

where, $\rho_{ij}(1 \le \rho_{ij} \le N)$ is the rank of *i*th gene in *j*th gene list. It can be noted that, for *i*th gene, $R_j^{(i)}\left(N^{-1} \approx 0 \le R_j^{(i)} \le 1\right)$ is a random variable (rv) [Farooqi *et al.* (2013)]. So, without loss of generality, another $rvr_j^{(i)}$ can be defined as

$$r_j^{(i)} = R_j^{(i)} - Q_2 \tag{6}$$

Where, Q_2 is second quartile (median) of rank scores, *i.e.* $R_j^{(i)}$. It may be noted that the rank scores of genes in each gene list is symmetrically distributed around Q_2(i.e. 0.5) (as rank scores are function of gene ranks). Further, to select informative genes, the following hypothesis was tested for each gene in Ω successively.

H_0 : *i*th gene is not informative for a given condition/ trait, *i.e.* $W_i \le 0$.

H_1 : *i*th gene is informative for a given condition/ trait, *i.e.* $W_i > 0$.

Where, W_i is the median deviated expected rank score for *i*th gene over all possible bootstrap samples. The bootstrap procedure coupled in the subject sampling model was used to ensure the iid assumptions of the rank scores [Ahmad *et al.* (2012)]. The test statistics for testing the above hypothesis was obtained as

Let for gene i, the r_j's be arranged in ascending order of their magnitude. Subsequently, the ranks 1, 2,

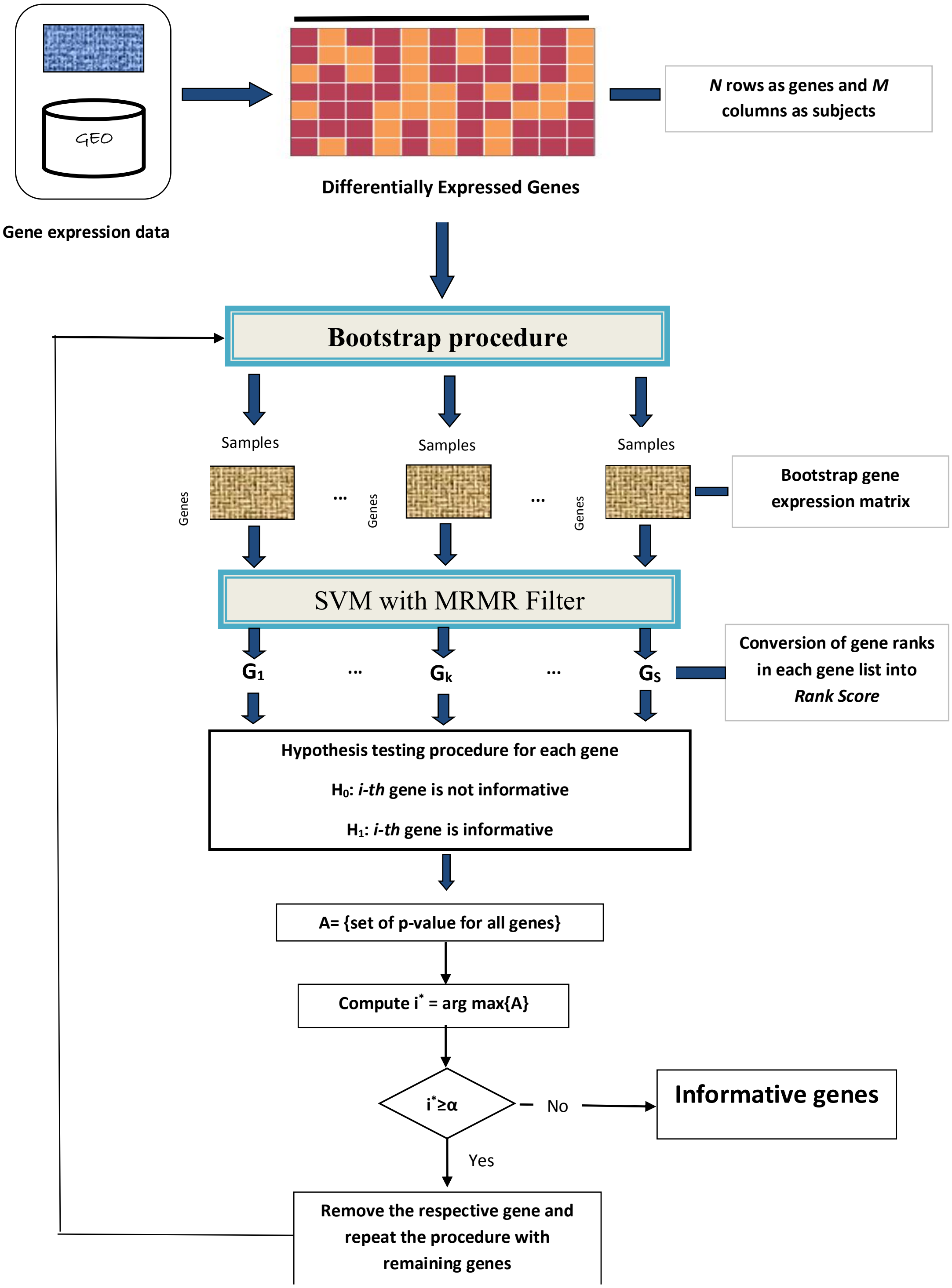
GEO
Gene expression data
Differentially Expressed Genes
N rows as genes and *M* columns as subjects
Bootstrap procedure
Samples
Samples
Samples
Genes
...
Genes
...
Genes
Bootstrap gene expression matrix
SVM with MRMR Filter
G_1
...
G_k
...
G_S
Conversion of gene ranks in each gene list into *Rank Score*
Hypothesis testing procedure for each gene
H_0: *i-th* gene is not informative
H_1: *i-th* gene is informative
A= {set of p-value for all genes}
Compute i^* = arg max{A}
$i^* \geq \alpha$
No
Informative genes
Yes
Remove the respective gene and repeat the procedure with remaining genes

Fig. 1: Schematic workflow of the developed algorithm

… S are assigned, keeping in mind their original signs. Let, A^+ be sum of the ranks of positive r_j's and A^- be the sum of the ranks of negative r_j's. Thus, for the computation of distribution of A^+, another rv (B_l) is defined as

$$B_{(l)} = \begin{cases} 1, \; if\ the\ |r_j|\ has\ rank\ l(>0) \\ 0, \; else \end{cases} \qquad (7)$$

Now the variables $B_{(l)}$ is independent Bernoulli variates and its mean and variance can be obtained as

$$E\left(B_{(l)}\right)=\frac{S}{2}\begin{pmatrix} S-1 \\ l-1 \end{pmatrix} B(l,\, S-l+1) \tag{8}$$

$$Var\left(B_{(l)}\right)=E\left\{B_{(l)}\left(1-E\left(B_{(l)}\right)\right)\right\} \tag{9}$$

Further, the mean and variance of the test statistic (A^+) can be obtained as

$$E\left(A^+\right)=\sum_{l=1}^{S} lE\left(B_{(l)}\right) \tag{10}$$

$$Var\left(A^+\right)=\frac{1}{4}\sum_{l=1}^{S} l^2\left\{E\left(B_{(l)}\right)\left(1-E\left(B_{(l)}\right)\right)\right\} \tag{11}$$

Under the null hypothesis H_0 : W_i = 0, the above equations can be expressed as

$$E_{H_0}\left(A^+\right)=\frac{1}{2}\sum_{l=1}^{S} l=\frac{S(S+1)}{4} \tag{12}$$

$$Var_{H_0}\left(A^+\right)=\frac{1}{4}\sum_{l=1}^{S} l^2=\frac{S(S+1)(2S+1)}{24} \tag{13}$$

As the number of bootstrap samples are quite large (S = 200), then under Lindeberg's central limit theorem [Rohatgi and Saleh (2015)], the test statistic (A^+) follow normal distribution asymptotically, *i.e.*

$$\frac{A^+ - E_{H_0}\left(A^+\right)}{\sqrt{Var_{H_0}\left(A^+\right)}} \sim N(0,1) \tag{14}$$

Based on the above test statistic, the statistical significance value for *i*th gene (p_i-value) was computed and accordingly the *p*-values of all the genes are obtained. Based on the *p*-value (less than 0.001), informative genes are identified.

2.7 Performance Evaluation Methods

Leave one out cross validation technique (loocv) has been used for the training of the classification model. Further, a confusion matrix was used to describe the performance of a classification model on a set of test data for which the true values are known. It gives illustration of the performance of an algorithm. Sensitivity, specificity, accuracy, PPV (Positive Predictive Value) and NPV (Negative Predictive Value) were the evaluation measures used to compute the performance of the model of binary classification.

3. Results and Discussion

3.1 Performance analysis based on selection criterion

The evaluation of developed method was carried out by using rice expression data of abiotic stress (salinity). Developed methodology consists of a test where null hypothesis is *i*th gene is not informative in nature (*i* = 1, 2, …, *N*, where *N* is the total number of genes) and alternate hypothesis is that the *i*th gene is informative gene. In order to check the feasibility of this test, we have constructed a graph between p-values of test statistics and the associated genes. We compare this criterion for selection of informative genes with other criterion used in other methods such as Boot-MRMR and MRMR. The graphical presentation of genes along with corresponding p-values for Boot-MRMR and proposed technique are shown in Figs. 2a and 2c, respectively. The distribution of MRMR weight (criterion used to select the informative genes) of the genes from MRMR technique is shown in Fig. 2b. It can be easily visualized from the Fig. 2c that the number of informative genes (genes with less p-values) selected from developed methodology are less as compared to the Boot-MRMR technique, which is the natural/desired case. The statistical significance values (p-values) of the genes are well separated in case of developed methodology. Therefore, in case of developed approach, a cut off p-value can be used more precisely to select informative genes in comparison to boot-MRMR approach. Thus, on the basis of comparative analysis of these graphs, we can say that developed methodology comparatively performs better than the existing technologies *viz*. Boot-MRMR and MRMR. Moreover, criterion for informative gene selection on the basis of p-values seems to be more statistically sound and meaningful as compared to other criterion of informative gene selections.

3.2 Performance analysis based on evaluation measures

Performance of the proposed methodology can be assessed through the classification ability of the selected informative genes. These selected genes were used to train and test the support vector machine classification model using linear kernel.

In order to test the model using these informative genes and training data, Leave- one-out cross-validation technique was used as it better performs when the

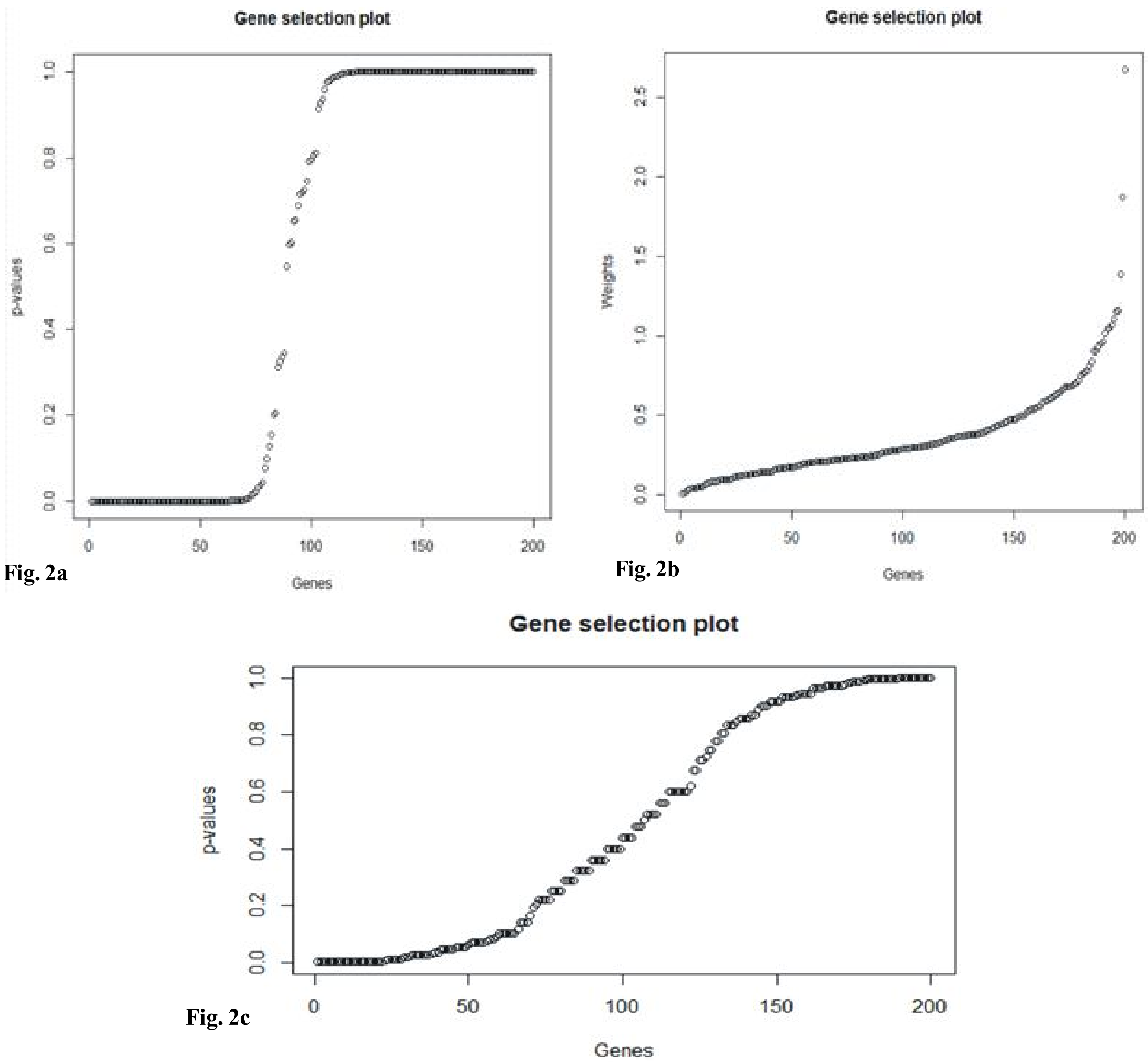

Fig. 2a

Fig. 2b

Fig. 2c

Fig. 2: Gene selection plot of (2a) Boot-MRMR (2b) MRMR (2c) Developed algorithm

Table 1: Confusion Matrix.

Evaluation measures	Developed	BOOT-MRMR	MRMR	SVM-RFE
Classification accuracy	0.9500	0.925	0.925	0.925
Sensitivity	0.9000	0.8500	0.8500	0.8500
Specificity	1.0000	1.0000	1.0000	1.0000
PPV	1.000	1.000	1.000	1.000
NPV	0.9091	0.8696	0.8696	0.8696

number of samples to train the model is less. The classification accuracy along with other evaluation measures such as sensitivity, specificity, PPV and NPV for informative gene selection techniques using rice gene expression data under salt stress are given in Table 1.

Classification accuracy of developed methodology was found to be 0.95 whereas for other methods, it is 0.925. Consequently, improved classification accuracy of proposed methodology has been observed which is higher than that of other gene selection techniques like Boot-MRMR, MRMR and SVM-RFE. This result indicates that genes selected by developed approach are more informative as compared to other techniques.

4. IGST R Software Package

To facilitate the use of proposed informative gene selection approach, we have developed an R software package, which includes IGST R package accompanying documentation and model real data example. This package can be freely downloaded from https://cran.r-project.org/web/packages/IGST. This software is capable of computing weights and p-values for genes using Bootstrap based resampling procedure. It also able to identify group of informative genes of given size based on the proposed approaches.

Acknowledgements

Authors sincerely acknowledge the computational and financial support provided by ICAR Network project on Agricultural Bioinformatics and Computational Biology (CABin Scheme), ICAR-Indian Agricultural Statistics Research Institute, New Delhi and Post Graduate School, ICAR-Indian Agricultural Research Institute, New Delhi.

References

Ahmad, T., A. Rai and P.M. Sahoo (2012). Comparison of bootstrap techniques for estimation of ratio in complexsurveys. *Int. J. Agricult. Stat. Sci.*, **8(1)**, 355-365.

Bolstad, B., R. Irizarry, M. Astrand and T. Speed (2003). A comparison of normalization methods for high density oligonucleotide array data based on bias and variance. *Bioinformatics*, **19(2)**, 185-193.

Dash, M. and H. Liu (1997). Feature selection for classification. *Intelligent Data Analysis*, **1(1-4)**, 131-156.

Dasgupta, P., T. Ahmad, A. Biswas and A. Rai (2018). A dual frame approach for estimating finite population total using ranked set sampling. *Int. J. Agricult. Stat. Sci.*, **14(1)**, 409-418.

Das, S., P.K. Meher, U.K. Pradhan and A.K. Paul (2017). Inferring gene regulatory networks using Kendall's tau correlation coefficient and identification of salinity stress responsive genes in rice. *Current Science*, **112(6)**, 1257.

Das, S., P.K. Meher, A. Rai, L.M. Bhar and B.N. Mandal (2017). Statistical approaches for gene selection, hub gene identification and module interaction in gene co-expression network analysis: An application to aluminum stress in soybean (*Glycine max* L.). *PLoS One*, **12(1)**, e0169605.

Das, S., A. Rai, D.C. Mishra and S.N. Rai (2018). Statistical approach for selection of biologically informative genes. *Gene*, **655**, 71-83.

Diaz-Uriarte, R. (2007). Gene SrF and varSelRF: a web-based tool and R package for gene selection and classification using random forest. *BMC Bioinformatics*, **8(1)**, 328. DOI:10.1186/1471-2105-8-328

Díaz-Uriarte, R. and S.A. De Andres (2006). Gene selection and classification of microarray data using random forest. *BMC Bioinformatics*, **7(1)**, **3**.

Ding, C. and H. Peng (2005). Minimum redundancy feature selection from microarray gene expression data. *J. Bioinformatics and Computational Biology*, **3(2)**, 185-205.

Efron, B. and R.J. Tibshirani (1994). *An Introduction to The Bootstrap*. CRC press.

Farooqi, M.S., R.K. Sanjukta, N. Sharma, A. Rai, D.C. Mishra, D.P. Singh and K.K. Chaturvedi (2013). Statistical and computational methods for detection of synonymous codon usage pattern s and gene expression. *Int. J. Agricult. Stat. Sci.*, **9**, 303-310.

Gautier, L., L. Cope, B.M. Bolstad and R.A. Irizarry (2004). Affy-analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics*, 20(3), 307-315.

Gentleman, R.C., V.J. Carey, D.M. Bates, B. Bolstad, M. Dettling, S. Dudoit, B. Ellis, L. Gautier, Y. Ge and J. Gentry (2004). Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.*, **5(10)**, 1-16.

Goeman, J.J. and P. Bühlmann (2007). Analyzing gene expression data in terms of gene sets: Methodological issues. *Bioinformatics*, **23(8)**, 980-987.

Guyon, I. and A. Elisseeff (2003). An introduction to variable and feature selection. *J. Machine Learning Research*, **3**, 1157-1182.

Guyon, I., J. Weston, S. Barnhill and V. Vapnik (2002). Gene selection for cancer classification using support vector machines. *Machine Learning*, **46(1-3)**, 389-422.

Liang, Y., F. Zhang, J. Wang, T. Joshi, Y. Wang and D. Xu (2011). Prediction of drought resistant genes in *Arabidopsis thaliana* using SVM-RFE. *PLoS One*, **6(7)**, e21750.

Mishra, D.C., S. Kumar, S.B. Lal, A. Saha, K.K. Chaturvedi and N. Budhlakoti (2018). TAGPT: A Web Server for Prediction of Trait Associated Genes using Gene Expression Data. *Annals of Genetics and Genetic Disorder.*, **1(1)**, 1003.

Mrinmoy, R., R. Anil, K.N. Singh and V.I. Ramasubramanian (2017). Modeling and forecasting of hybrid rice yield using a grey model improved by the genetic algorithm. *Int. J. Agricult. Stat. Sci.*, **13(2)**, 563-566.

Mundra, P.A. and J.C. Rajapakse (2009). SVM-RFE with MRMR filter for gene selection. *IEEE Transactions on Nanobioscience*, **9(1)**, 31-37.

Peng, H., F. Long and C. Ding (2005). Feature selection based on mutual information: criteria of max-dependency, max-relevance and min-redundancy. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, **8**, 1226-1238.

Pyingkodi, M. and R. Thangarajan (2017). Meta-analysis in autism gene expression dataset with biclustering methods using random cuckoo search algorithm. *Asian J. Res. Social Sciences and Humanities*, **7(2)**, 186-194.

Rohatgi, V.K. and A.M.E. Saleh (2015). *An Introduction to Probability and Statistics*. John Wiley & Sons.

Tang, Y., Y.Q. Zhang and Z. Huang (2007). Development of two-stage SVM-RFE gene selection strategy for microarray expression data analysis. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, **4(3)**, 365-381.

Vapnik, V. (1998). The support vector method of function estimation in Nonlinear modeling. Springer, Boston, MA.

Wang, J., L. Chen, Y. Wang, J. Zhang, Y. Liang and D. Xu (2013). A computational systems biology study for understanding salt tolerance mechanism in rice. *PLoS One*, **8(6)**, e64929.

Yu, L. and H. Liu (2004). Efficient feature selection via analysis of relevance and redundancy. *J. Machine Learning Research*, **5**, 1205-1224.

Zhu, S., D. Wang, K. Yu, T. Lia and Y. Gong (2010). Feature selection for gene expression using model-based entropy. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, **7(1)**, 25-36.