IJCSE International Journal of Computer Sciences and Engineering Open Access

Research Paper Vol.-6, Issue-11, Nov 2018 E-ISSN: 2347-2693

Cross Validation Of Supervised Machine Learning Models Based On Random Forest and Support Vector Machine Techniques for 12S rRNA Molecular Marker: Implementation, Comparison and Utility

Rameshwar Pati[1, 2*], Ajey Kumar Pathak[1], Navita Srivastava[2]

[1]ICAR-National Bureau of Fish Genetic Resources, Lucknow- 226002, (U.P.) India.
[2]Dept. of Computer Science, Awadhesh Pratap Singh University, Rewa-486003 (M.P.), India.

**Corresponding Author: rameshwarpati@gmail.com, Tel: 9415854020*

Available online at: www.ijcseonline.org

Accepted: 20/Nov/2018, Published: 30/Nov/2018

Abstract- Folding plays imperative role in the cross validation studies of machine learning based models. The folding divides the original sample into training and test sets, which evaluate performance of the machine learning based models and present scenarios for optimising the efficacy of such models. The present study discusses about the computational approaches applied for preparing training and test sets at different folds from 12S rRNA molecular marker sequence dataset of fish and application of these sets to estimate the performance of the proposed models based on machine learning techniques viz. Random Forest and Support Vector Machine. Additionally, the study presents the comparative accounts on efficacies of these models estimated at different folding. The findings from the study showed that folding has linear relationship with the efficacy of the model. The model with random forest was found better for solving the classification problems of the molecular marker sequence data. This study provides understanding on utility of the folding level in increasing the efficacy of the machine learning based methods and suggests for suitable machine learning method for solving the multiclass problem data especially where the identification using the molecular markers sequence data is involved.

Accessibility of the supporting documents-- http://mail.nbfgr.res.in/FishIdMar/.

Keywords-- Machine learning method, Random forest, Support vector machine, Folding level, 12S rRNA, Cross validation.

I. INTRODUCTION

Machine learning featuring data analysis automates the analytical model building [1] and make such models capable to find hidden insights without being explicitly programmed. Scientific community concentrating on new techniques and methods that would reduce the processing time and increase the accuracy level at the same time [2]. Processing of raw data increases at respected domain due to the advancement technique in data mining [3]. All machine learning classifiers improve with experience during the cross-validation phase. Crossvalidation plays an imperative role in estimating the performance of a classifier and number of folds during cross-validation accounts much in providing the training and test datasets and their subsets on which machine learning classifier works. For example in 10-fold cross validation, data is broken into 10 sets of size n/10. Now train classifier on 9 datasets and test on 1. Repeat 10 times and take a mean accuracy.

Due to innovation in new technological development of high-throughput sequencing and further reduction in sequencing cost, an enormous amount of molecular data are generated and stored in public repositories to make them available for researchers. Only the few workers used the sequence data of different molecular markers stored in public repositories for developing the machine learning models based on supervised or unsupervised methods to solve the classification problems. Random forest and Support vector machine classifiers are widely used supervised learning methods to solve classification problems and used extensively to solve the classification problems using molecular marker sequence data [4]. Molecular data of ribosomal RNA (rRNA) plays an important role in structure prediction and protein synthesis [5], [6]. Among vertebrates, sequences of 12S ribosomal RNA (12S rRNA) gene have been used widely for phylogenetic study among different levels of taxa such as families, genera, and species [7], [8], [9]. Because of the high mutation rate, 12S rRNA gene is used for species identification as it produces a significant amount of sequence variation in closely related species. In machine learning, preparation of datasets for 12S rRNA sequence data at different folding levels produces different conservation of sequence variation. A huge amount of 12S rRNA gene raw data is available in the public domain. Several methods have been proposed to analyse 12S rRNA data like RAPD fingerprinting [10], DNA hybridization [11] restriction fragment length polymorphism [12] and real-time

PCR [13]. Till now, hardly any worker has computationally analysed the 12S rRNA data for identification of fish using the supervised machine learning classifiers at different foldings.

The present work discusses the cross-validation studies of 12S rRNA data based on two supervised learning models constructed using the Random Forest (RF) and Support Vector Machine (SVM) techniques at 3 and 5 folds. Additionally, collection of raw data from public domain and its curation with feature extraction and finally developed a model are explain in material and method section of the paper. Finally in result section, work presents the comapartive accounts on the performance of both classifiers estimated for solving the identification problem using 12S rRNA gene sequence dataset of fish species.

II. MATERIAL & METHOD

A. Primary dataset:

National Centre for Biotechnology Information (NCBI) [14] was used for downloading raw sequences of 12S rRNA of six fish taxa viz. Coelacanthimorpha, Hyperoartia, Hyperotreti, Actinopterygii, Dipnoi, and Chondrichthyes using the query “("Coelacanthimorpha" OR "Hyperoartia" OR "Hyperotreti" OR "Actinopterygii" OR "Dipnoi" OR "Chondrichthyes") AND 12S rRNA” in the nucleotide search option of NCBI. In totality, 25729 records of 12S rRNA confirming to 687 MB size were downloaded on 24 March 2017.

B. Feature extraction and conversion

The downloaded raw sequences of 12S rRNA are combination of nucleotide sequences of four base pair of A, T, C and G (Adenine (A), Thymine (T), Guanine (G) and Cytosine (C). As machine learning classifier work on numerical feature vector [15], each nucleotide charcater from the nucleotide sequences was extracted and converted into the the numerical feature vector using the four-bit binary fixed mapping method. Thus, ‘A’ was represented by '1000', ‘T’ by '0100', 'G' by '0010' and 'C' by '0001' using the four-bit pattern. Other characters were represented by ‘1111’. To do this task an algorithm using the Perl script language under Windows platform was designed and implemented. This algorithm converts the string feature vectors into the tab spaced characters and then into a four-bit binary pattern. The reason for converting nucleotide sequences into binary feature vector was to acheve the better classification performance on the diverse datasets [16].

C. Preparation of datasets

12S rRNA molecular marker is a highly mutated gene and available almost in all vertebrateswith an average of 950 bp length [17]. In the present study, 650 bp of 12S rRNA was taken up for preparing training and test datasets at 3 and 5 folds. For preparing training and test datasets at different folds, three scripts ‘IndividualGeneSeqParse_Dowloaded.pl’, ‘SameSeqLengthMaker.pl’ and ‘TrainednTes Dataset.pl’ were designed and implemented using the Perl language under the Windows operating environment. These scripts are logically linked together and work as a pipeline in which output produced by one script works as input for another. The first Perl script (‘IndividualGeneSeqParse_Dowloaded.pl’) takes original sequence dataset of 12S rRNA in fasta format file as input and applies different type filters to remove the unwanted sequences based on the values input by the user. This script provides the ability to generate a dataset of sequences having sequence length from 300 to 1800 base pair. Thereafter, the second Perl scripts (‘SameSeqLengthMaker.pl’) applies a filter on the output of the previous script to prepare the sequence datasets of 650 bp length only. This output is a curated dataset which is used further as input by third script (‘TrainednTesDataset.pl’) that prepares the binary coded training and test datasets at different folds. In the present study 3 and 5 values were used as foldings to generate the training and test datasets.

D. Design of framework and model building

For cross-validation studies of 12S rRNA molecular marker at different folds, the machine learning models based on random forest and support vector machine algorithms were designed and implemented under R platform. These models accept training and test datasets as input. The training sets were used for training the models and test sets were used to estimate the performance of the models. Figure 1 shows the architecture of the framework built for identifying fish using 12S rRNA molecular marker sequence data based on RF and SVM techniques.. The framework also provides the accuracy assessments predicted by these models.

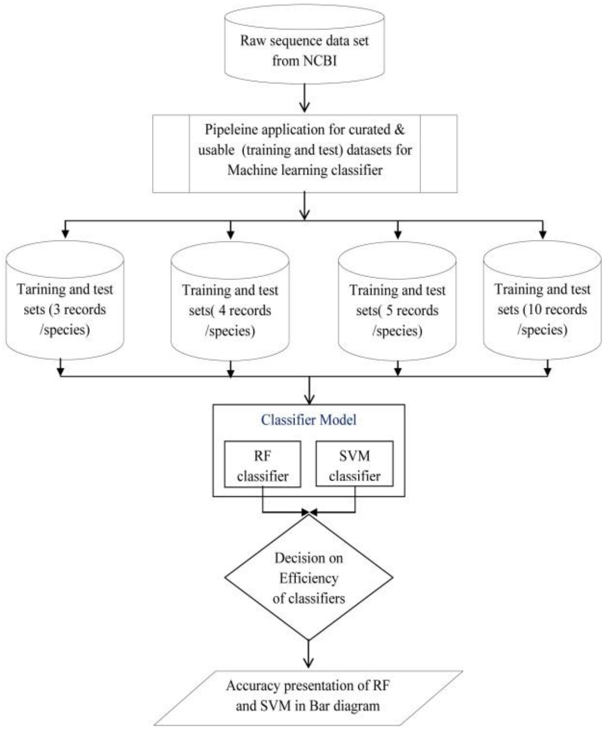

Raw sequence data set from NCBI
Pipeleine application for curated & usable (training and test) datasets for Machine learning classifier
Tarining and test sets (3 records /species)
Training and test sets(4 records /species)
Training and test sets(5 records /species)
Training and test sets (10 records /species)
Classifier Model
RF classifier
SVM classifier
Decision on Efficiency of classifiers
Accuracy presentation of RF and SVM in Bar diagram

Figure 1: Architecture of the framework for identying fish using 12S rRNA molecular marker sequence data.

III. RESULTS & DISCUSSION

A. Curated dataset

A total of 25729 sequences of 12S rRNA gene were downloaded using the query expression. A curated sequence dataset of 6956 sequences was obtained as output from 'IndividualGeneSeqParse_Dowloade d.pl' program and this dataset was further divided into four different datasets based on number of records per species (3, 4, 5 and 10). In order to obtain the curated sequence dataset of uniform length, 'SameSeqLengthMaker.pl' program was applied on the datasets produced as output by the earlier program and in this way the sequence datasets of 650 base pair length were obtained, which were used further for feature extraction and conversion. Finally, 'TrainednTesDataset.pl' program was run with 3 and 5 folds on the outputs produced by the 'SameSeqLengthMaker.pl'. This program provided binary coded training and test sets as output in respect of each dataset. In three folding, 3 subsamples of training and test datasets were produced whereas in five folding the 5 subsamples of training and test datasets were generated. Thus, a total of 17 subsamples for training and 17 subsamples for test datasets were generated based on the curated records in the four datasets. Table 1 presents the details on the four curated datasets accomplished after applying 'IndividualGeneSeqParse_Dowloaded.pl' program and Table 2 lists the training and test datasets generated using 3 and 5 folds as output produced by 'TrainednTesDataset.pl' program.

Table1: List of curated datasets based on number of sequence records per species.

Marker	Sequence frequency	Number of species	Curated datasets
12S rRNA	3	100	2849
	4	100	2590
	5	100	2433
	10	40	2010

Table2: List of training and test datasets at different folding level.

Folding Level	Frequency of the records	Test Datasets		Training Datasets	
		Datasets	No. of records	Datasets	No. of records
3	3	I	240	I	2609
		II	240	II	2609
		III	240	III	2609
	4	I	150	I	2440
		II	150	II	2440
		III	150	III	2440
	5	I	110	I	2323
		II	110	II	2323
		III	110	III	2323
	10	I	43	I	1967
		II	43	II	1967
		III	43	III	1967
5	10	I	43	I	1967
		II	43	II	1967
		III	43	III	1967
		IV	43	IV	1967
		V	43	V	1967
Total number of Datasets:		**17**		**17**	

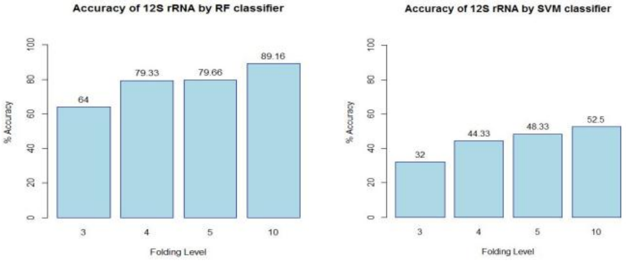
B. Availability and accessibility of models and data

An online downloadable FTP folder named as FishIdMar accessible at URL: http://mail.nbfgr.res.in/FishIdMar/. was created to facilitate the scientific community for making use of the model and data both. It is freely available for all type of user and in this folder the prepared datasets and FishIdMar model both has been provided to validate the work at different folding level and to use in their works. The nucleotide sequences datasets of fish species are available in three and five levels both having minimum 3 or 4 or 5 or 10 records per species. These datasets have already been validated by FishIdMar model based on random forest and support vector machine techniques.

C. Analysis of the dataset

The proposed models based on RF and SVM were constructed on R platform to identify fish using the training and test sets obatined at 3 and 5 foldings from the original sequence datasets of different genes. Figure 2 presents the accuracies estimated by these proposed models. From this figure, it is evident folding has linear relationship with the accuracy of both the classifiers and follows the uniform trend. In both the folding situations, the performance of RF was estimated better than SVM.

Though several studies in the past have been done to validate the developed models, hardly any study provides information about increasing the efficacy of the model prepared for the molecular marker sequence datasets. In the present study, the cross-validation studies of the proposed models were done using the training and test datasets generated at 3 and 5 foldings and based on the number of records per species. This work justifies the work done by earlier workers [18], [19], [20], [21]. Further earlier works were performed on the small datasets of different organism available online [18], [19], [20], [21] while this study not only develops the indigenous datasets of different fish taxa from the open source of NCBI at the different folding level of fixed base pair length, but also higlights on the use of these datasets by the constructed models and assessing the predicted efficacies. Besides, these dataset are directly usable to any machine learning classifier.



(A) **(B)**

Figure 2: Efficiency of four datasets at different folding level by **(A)** RF and **(B)** SVM

IV. CONCLUSION

The present study presents a case study in which performance of the constructed model was evaluated at different folding levels for 12S rRNA molecular marker dataset. The study also determines the best classifier for identification of fish and model based on RF was estimated as the best classifier relative to SVM. In addition, this study highlights the primitive role of foldings in getting the usable datasets and determing the performance of the machine learning models in relation to foldings. During the study, it was found that the selection of suitable folding level and base pair length play an effective role in accomplishing the better accuracy. Thus, this study can play a pivotal role for the scientific community interested in machine learning based models who wish to estimate the performance of their constructed models and assess the efficacies in relation to others.

ACKNOWLEDGEMENT

The authors are thankful to the Director of ICAR-NBFGR, Lucknow, Ministry of Agriculture and Farmer Welfare, Govt. of India, New Delhi and Department of Computer Science, Awadhesh Pratap Singh University, Rewa for providing the necessary facilities to carry out the work.

Author disclosure

Authors disclose that this manuscript has not been submitted elsewhere for publication and it is their original work. There is no conflict of interest among authors.

REFERENCE

[1] T. Mitchell, "Machine Learning, McGraw Hill Publisher, New York, NY," pp-441, 1997.

[2] S.U. Bohra, P.V. Ingole , "Review on Neural Network Based Approach Towards English Handwritten Alphanumeric Characters Recognition", International Journal of Computer Sciences and Engineering, Vol.1, Issue.3, pp.22-25, 2013.

[3] V. Bhambri, "Data Mining as a Solution for Data Management in Banking Sector", International Journal of Computer Sciences and Engineering, Vol.1, Issue.1, pp.20-25, 2013.

[4] P. Yang, , Hwa Y. Yang, B. Zhou, and Y. Zomaya, et al., "A review of ensemble methods in bioinformatics," Current Bioinformatics, vol. 5(4), pp. 296–308, 2010.

[5] A.E. Dahlberg, "The functional role of ribosomal RNA in protein synthesis," Cell, vol. 57**,** pp. 525–529, 1989.

[6] H.F. Noller, "Structure of ribosomal RNA," Annual Review Biochemistry, vol. 53, pp. 119–162, 1984.

[7] K.M. Kjer, "Use of rRNA secondary structure in phylogenetic studies to identify homologous positions: an example of alignment and data presentation from the frogs," Molecular Phylogenetics and Evolution, vol. 4, pp. 314–330, 1995.

[8] A.M. Simons and R.L. Mayden, "Phylogenetic relationships of the western North American phoxinins (Actinopterygii: Cyprinidae) as inferred from mitochondrial 12S and 16S ribosomal RNA sequences," Molecular Phylogenetics and Evolution, vol. 9, pp. 308–329, 1998.

[9] J. Alves-Gomes, G. Orti, M. Haygood, W. Heiligenberg, and A. Meyer, "Phylogenetic analysis of South American electric fishes (order: Gymnotiformes) and the evolution of their electrogenic system: a synthesis based on morphology, electrophysiology, and mitochondrial sequence data," Molecular Biology and Evolution, vol. 12, pp. 298-318, 1995.

[10] J.C.I. Lee and J.G. Chang, "Random amplified polymorphic DNA polymerase chain reaction (RAPD PCR) fingerprints in forensic species identification," Forensic Science International, vol. 67(2), pp. 103–107, 1994.

[11] R.S. Blackett and P. Keim, "Big game species identification by deoxyribonucleic acid (DNA) probes," Journal of Forensic Sciences, vol. 37(2), pp. 590–596, 1992.

[12] R. Meyer, C. Höfelein, J. Lüthy and U. Candrian, "Polymerase chain reaction-restriction fragment length polymorphism analysis: a simple method for species identification in food," Journal of AOAC International, vol. 78(6), pp. 1542–1551, 1995.

[13] M.L. López-Andreo, Lugo, A. Garrido-Pertierra, M.I. Prieto and A. Puyet, "Identification and quantitation of species in complex DNA mixtures by real-time polymerase chain reaction," Analytical Biochemistry, vol. 339(1), pp. 73–82, 2005.

[14] NCBI Resource Coordinators, "Database resources of the National Center for Biotechnology Information," Nucleic Acids Research, vol. 44, pp. D7–D19, 2016.

[15] X. Zhang, J. Lee, and L.A. Chasin, "The effect of nonsense codons on splicing: a genomic analysis," RNA,vol. 9, pp. 637–639, 2006.

[16] C.M. Vander Walt and E. Barnard, "Data characteristics that determine classifier performance," Proceedings of the 17th Annual Symposium of the Pattern Recognition Association of South Africa, pp. 166-171, 2006.

[17] Li. Yang, Z. Tan, D. Wang, L. Xue, M. Guan, T. Huang, and R. Li, "Species identification through mitochondrial rRNA genetic analysis," Scientific Reports, vol. 4, pp. 4089, 2014.

[18] P.K. Meher, T.K. Sahu and A.R. Rao, "Identification of species based on DNA barcode using kmer feature vector and Random forest classifier," Gene, vol. 592(2), pp. 316-24, 2016.

[19] C. Guisande, A. Manjarrés-Hernández, P. Pelayo-Villamil, C. Granado-Lorencio, I. Riveiro, A. Acu˜na, E. Prieto-Piraquive, E. Janeiro, J.M. Matías, C. Patti, B. Patti, S. Mazzola, S. Jiménez, V. Duqueg and F. Salmerón, "IPez: An expert system for the taxonomic identification of fishes based on machine learning techniques," Fisheries Research, vol. 102, pp. 240–247, 2010.

[20] Satoh P. Takashi, Miya Masaki, Mabuchi Kohji and Nishida Mutsumi, "Structure and variation of the mitochondrial genome of fishes," BMC Genomics. Vol. 17,pp. 719, 2016.

[21] E. Weitschek, Iulia G. Fiscon and G. Felici "Supervised DNA Barcodes species classification: analysis, comparisons, and results," BioData Mining, 7, pp. 4, 2014.

Authors Profile

Rameshwar Pati received his M.Phil. degree in Computer Science from the Department of Computer Science and Engineering, Annamalai University, Annalalainager, TN in 2011. Currently, he is pursuing Ph.D. (Computer Science) from Awadhesh Pratap Singh University, Rewa, MP. His research interests include Database development, Big data Analysis and Machine learning.

Ajey Kumar Pathak received his Ph.D. (Computer Science and Information Technology) from the Department of Computer Science, Sam Higginbottom University of Agriculture, Technology and Sciences, Allahabad, UP in 2016. Currently Dr. Pathak is working as a Senior Scientist in ICAR-National Bureau of Fish Genetic Resources, Lucknow, UP. His research interests include Fish Bioinformatics, Fisheries Database Development, Machine learning, Application of Geographic Information System and Remote Sensing technologies in Fisheries Resource Assessment and Management.

Navita Srivastava received her Post Doc (Physics) from IUCAA, Pune in 1993. Currently Dr. Srivastava is a Professor in the Department of Computer Science at Awadhesh Pratap Singh University, Rewa, MP. Her research interests include Machine learning, Bioinformatics, Internet Security, A&A and Aeronomy.