EnsembleFS: an R toolkit and a web-based tool for a filter ensemble feature selection of molecular omics data

Polewko-Klim Aneta[1[0000−0003−1987−7374]], Grablis Paweł[1], and Rudnicki Witold R.[1,2,3[0000−0002−7928−4944]]

[1] Faculty of Computer Science, University of Bialystok, K. Ciołkowskiego 1M, 15-245, Poland
[2] Computational Center, University of Bialystok, K. Ciołkowskiego 1M, 15-245, Poland
[3] Interdisciplinary Centre for Mathematical and Computational Modelling, University of Warsaw, Kupiecka 32, 03-046, Warsaw, Poland

Abstract. The development of more complex biomarker selection protocols based on the machine learning (ML) approach, with additional processing of information from biological databases (DB), is important for the accelerated development of molecular diagnostics and therapy.
In this study, we present *EnsembleFS* user-friendly R toolkit (R package and Shiny web application) for heterogeneous ensemble feature selection (EFS) of molecular omics data that also supports users in the analysis and interpretation of the most relevant biomarkers. *EnsembleFS* is based on five feature filters (FF), namely, U-test, minimum redundancy maximum relevance (MRMR), Monte Carlo feature selection (MCFS), and multidimensional feature selection (MDFS) in 1D and 2D versions. It uses supervised ML methods to evaluate the quality of the set of selected features and retrieves the biological characteristics of biomarkers online from the nine DB, such as Gene Ontology, WikiPathways, and Human Protein Atlas. The functional modules to identify potential candidate biomarkers, evaluation, comparison, analysis, and visualization of model results make *EnsembleFS* a useful tool for selection, random forest (RF) binary classification, and comprehensive biomarker analysis.

Keywords: ensemble feature selection · machine learning · omics

1 Introduction

The molecular omics data are generally unbalanced and high-dimensional with a low sample size, and have complex correlation structures. Although multiple bioinformatic tools have been developed to analyze omics data, the practical process of selecting, evaluating, and analyzing crucial biomarkers from these data is a significant challenge for researchers.

Various feature selection (FS) methods implemented in the R and Python packages are usually used to construct the computational pipeline for the discovery of biomarkers from omics data. Researchers often use open-source software

ICCS Camera Ready Version 2024
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-63772-8_7

for biomarker identification available in public repositories, such as GitHub, as well as non-commercial automated software, such as, for example, the OmicSelector [18]. However, the use of these tools usually requires a certain level of experience in programming and statistics, knowledge of ML methods, and specific hardware resources. Moreover, these ready-made FS procedures are usually designed and optimized for specific types of omics data [3] and particular research tasks. Only a handful of tools are specialized in selecting biomarker candidates from omics datasets for supervised ML methods. In the literature, we found only a few tools that partially address this issue, such as the FeatureSelect software in MATLAB [13], the OmicSelector tools based on deep learning [18], the standalone program BioDiscML [11], MRMD3.0 Python tool [8], and the mixOmics R package [17]. FeatureSelect uses three classes of FS methods and then applies optimization algorithms to find the optimal feature subset and create predictive models. MRMD3.0 uses seven FF, three wrapper methods, and seven embedding methods to search for the best features for classifiers. BioDiscML provides multiple multivariate data analyses and uses wrappers to find the optimal combination of features to predict outcomes. The MixOmics offers multivariate FS methods for exploring and integrating biological data sets. The software libraries mentioned above focus on FS and ML techniques that can find the minimal and optimal combination of features for predicting models or multi-omics data integration tasks. These tools do not have functionalities that allow the user to retrieve biological information about biomarkers from the database and use methods that are susceptible to noise and instability.

Here, we propose a comprehensive tool for biomarker discovery in quantitative omics data that allows users to: (i) select the top biomarker candidates using either an ensemble of FS methods or any individual FS method, (ii) build and evaluate predictive models using the RF classifier, (iii) evaluate the quality of the feature set, (iv) benchmark model results for various FS methods, (v) retrieve biological information about the top biomarkers from the nine biological DB, e.g molecular function, cellular component, and biological process from the Gene Ontology (GO) [2], signalling pathways from the Kyoto Encyclopedia of Genes and Genomes (KEGG) [15], and disease phenotypes from the Human Phenotype Ontology (HP) [10].

2 Methods

2.1 Feature selection and classification algorithms

The *EnsembleFS* uses the U-test, MRMR [4], MCFS [5], MDFS-1D and MDFS-2D methods [14] to remove irrelevant variables. These feature filters are not related to the classifier and have better generalization properties than wrappers and embedded methods. [9].

The MDFS measures the decrease in the information entropy of the decision variable due to the knowledge of k-dimensional tuples of variables and measures the influence of each variable in the tuple [14]. This FF performs performs an exhaustive search of all possible k-tuples and assign to each variable a maximal

ICCS Camera Ready Version 2024
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-63772-8_7

information gain due to a given variable that was achieved in any of the k-tuples that included this variable. The 2D version of this algorithm (MDFS-2D) can capture synergistic interactions between feature pairs and the decision variable.

The MRMR method is based on mutual information (MI) as a measure of the relevancy and redundancy of features, where the feature redundancy is an aggregate MI measure between each pair of features in the selected feature subset, and relevance to a class variable is an aggregate MI measure between each feature with respect to the class variable.

The MCFS method is based on a Monte Carlo approach to select informative features. This algorithm is capable of incorporating interdependencies between features. The MCFS offers several cutoff methods (e.g. critical angle, k-means, and permutations) for discerning informative and non-informative features.

The random forest algorithm [1] was used to construct predictive models. This algorithm works well in data sets with a small number of objects, has few tuneable parameters that do not relate directly to the data, is very rarely faulty, and usually gives results that are often the best or very close to the best results achievable by any classification algorithm [6].

Algorithm 1: EFS$(l, f, S = \{P_1, \ldots, P_k\})$ the ensemble FS algorithm with RF classifier (l).

input : Feature filters f_j, $j = 1, \ldots, m$
Dataset $S = \{(y, X)\}$ with n entries of p features $V = \{v_1, \ldots, v_p\}$ belonging to one of two classes, randomly split into k partitions P_i
output: Combined set of informative features F
Ranked informative feature set F_j, $j = 1, \ldots, m$
Performance estimation metric E_j, $j = 1, \ldots, m$
Feature selection stability measure A_j, $j = 1, \ldots, m$

repeat r **times**
  foreach S_i **do**
    Generate the training set $S_{\setminus i}(V) \leftarrow S(V) \setminus P_i(\mathrm{V})$
    foreach f_j **do**
      Perform feature selection on the training set $W_i \leftarrow f(S_{\setminus i}(V))$
      Collect the ranked informative feature set $W_i = \{v_1, \ldots, v_d\}$
      Remove highly correlated features with W_i
      Build the model on the training set $L_i \leftarrow l(S_{\setminus i}(U_i))$ using top N features U_i with W_i
      Performance estimation E_i: use the model L_i on a test set P_i
    end
  end
end
$E_j \leftarrow \frac{1}{r \cdot k} \Sigma E_i$, $i = 1, \ldots, r \cdot k$
Assess the FS stability A_j of $r \cdot k$ feature sets U_i
Collect the feature set F_j from $r \cdot k$ sets U_i by using the majority voting strategy, for each of m feature filters
Collect combined feature set $F = \bigcup_{j=1}^{m} F_j$ or $F = \bigcap_{j=1}^{m} F_j$

ICCS Camera Ready Version 2024
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-63772-8_7

2.2 Ensemble feature selection

The process of selecting relevant features from the original dataset and the model-building procedure executed in the *EnsembleFS* is shown in Algorithm 1. The area under the receiver operator curve (AUC), the accuracy (ACC) and the Matthews correlation coefficient (MCC) are used to assess the performance of the model, and the Lustgarten stability measure (ASM) was used [12] to assess the stability of selection.

3 *EnsembleFS* an R toolkit

EnsembleFS is based on carefully chosen statistical and ML methods recommended for biomedical data and uses feature filters based on alternative approaches: statistical, information theory, and methods sensitive to interactions between variables. It allows the user to select and rank relevant biomarkers from quantitative omics data using an ensemble of various FS algorithms (U-test, MRMR, MCFS, MDFS-1D, and MDFS-2D). The user can modify the list of FS methods used and the values of their parameters. The quality of the feature set can be verified by applying the RF classification algorithm within a stratified k-fold cross-validation (CV) or alternatively within size-k random sampling, repeated n times. Redundant and correlated features can be removed from extracted feature subsets. The stability of the feature sets returned by *EnsembleFS* is measured using ASM, while the performance of the predictive models is estimated using AUC, ACC, and MCC metrics. *EnsembleFS* allows to set the values of selected parameters for ML models, such as the top N number of features that the classifier will use. It also allows the user to compare the results of the predictive models obtained using the features returned by different FS methods. The results of the ML models are visualized in the form of interactive tables and plots. The final feature set for each filter is obtained by majority voting on the CV results. The combined set of biomarkers for biological analysis is obtained, depending on the user choice, as the intersection or union of the results of all the filters used.

EnsembleFS accepts data that include different biomarker identifiers (ID), such as Ensembl gene ID, NCBI Entrez gene ID, and Uniprot IDs. It should be underlined that *EnsembleFS* allows the generation of final report files that include the results of ML models and crucial information on the top genes from nine biological DBs.

3.1 Web application

EnsembleFS web application (app) has an interactive web interface for data analysis and visualization. This tool consists of a module for selecting relevant biomarkers and a module for collecting biological information on genes. The functionalities of these modules are available via *Feature Selection* tab and *Gene Information* tab, respectively. The general functional specification of these basic software modules is presented in Figure 1. *EnsembleFS* web app is available

ICCS Camera Ready Version 2024
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-63772-8_7

online at https://uco.uwb.edu.pl/apps/EnsembleFS (webserver demo). It is open source, free software under an MIT license. *EnsembleFS* web app architecture, the source code, workflow, tutorial and the exemplary report of feature selection and modelling results are described in detail in the *Home* tab, *Help* tab, and project home page https://github.com/biocsuwb/EnsembleFS.

3.2 R package

EnsembleFS R package includes software to select, collect, analyze and interpret the top biomarkers with omics data. Compared to the *EnsembleFS* web app, the R package includes the *ensembleFS()* dynamic function that allows the user to easily add any other FS method to the default list of five basic FF (*methods* input parameter). Users can manually set all the hyperparameters of the model. The size of input data is limited only by the computer's performance. Source code, examples, implementation details, and documentation are available on GitHub (https://github.com/biocsuwb/EnsembleFS-package).

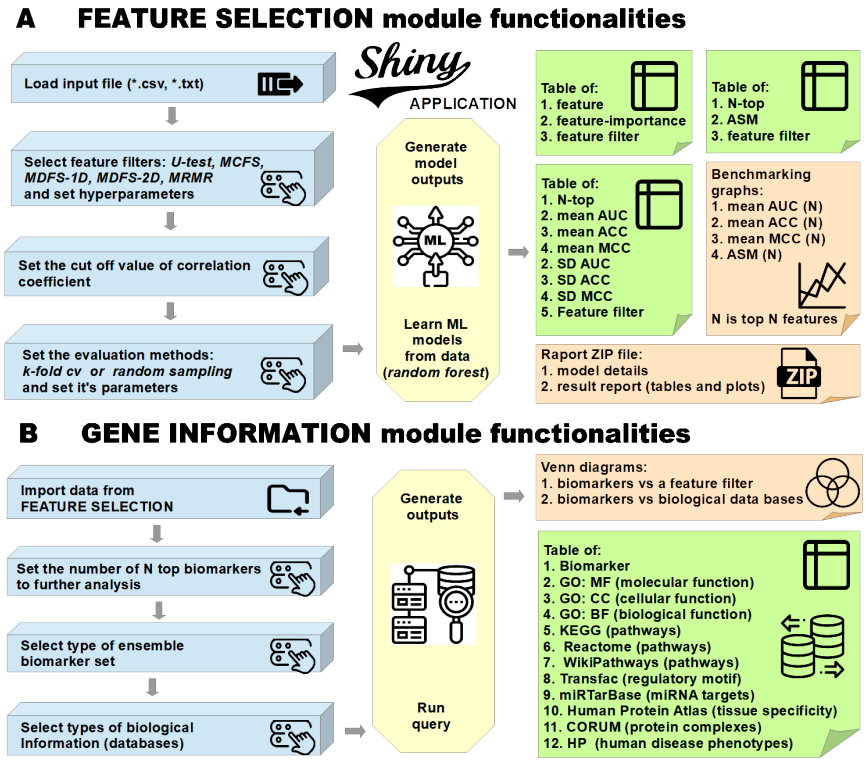

Fig. 1. Main functionality modules of *EnsembleFS* web app: A) Feature Selection tab, B) Gene information tab. Cuboids represent the interaction between *EnsembleFS* and the user, and the octagons represent *EnsembleFS* processes. For notes, see text.

ICCS Camera Ready Version 2024
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-63772-8_7

4 Use case

To demonstrate selected capabilities of *EnsembleFS* web app in a real case study, we used RNA-seq data from the TCGA-LUAD (`https://www.cancer.gov/tcga`) program [7]. The description of the data set, the preprocessing procedure, and the example results of feature selection and classification of tumor vs normal tissue are included in [16] and the Help tab → Example sub-tab. For testing purposes, we used only 574 samples and 2000 differentially expressed genes (DEGs) with the highest difference in gene expression level. To find the most relevant DEGs for the classification of tissue types, we performed the ensemble FS. Default parameters were used for all FS methods. The 0.3 random sampling, with 10 iterations, was selected for model validation. We conducted the quantitative analysis of the most informative DEGs and compared the performance of the prediction models for each FS method with the top N features. Our analysis shows that the five FS methods identified 1608 unique DEGs in total. MDFS-2D filter identified the highest number of relevant features (1024 DEGs) in ten subsets of features. The best predictive model (ACC = 0.996 ± 0.016) was obtained with the top 30 features selected by the MRMR method. The best overall predictive results were achieved by the RF classifier with at least the top 75 features for all filters (Fig.2 (left panel)). In this regard, the DEGs returned by individual FS algorithms for N = 75 were chosen for further biological analysis. It should be noted that the final DEG sets selected by different filters for N = 75 were quite divergent (Fig. 2(right panel)). Although 136 DEGs were selected in total, none of the DEGs was identified by all FS methods.

In the next step, we submitted queries to the nine biological DBs and collected information on the 51 genes found. Molecular function categories for 22 genes were indicated. The fifty-nine GO biological processes were examined for 23 genes. Nineteen significant cellular components were found for 12 genes.

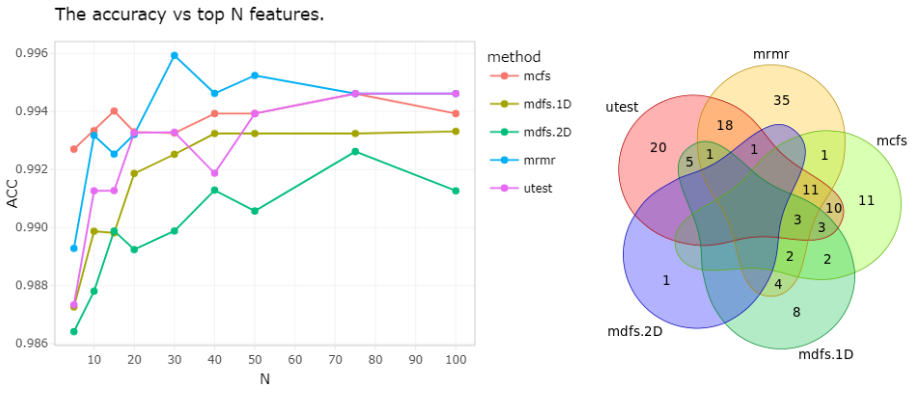

Fig. 2. Left panel: the average values for the ACC between 10 feature subsets as a function of N top features for all filters for LUAD data. Right panel: the final number of the most relevant biomarkers with FS methods for N = 75. See notation in text.

ICCS Camera Ready Version 2024
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-63772-8_7

Table 1. Execution times (hh:mm:ss) for a single iteration of the FS algorithm and RF classification for the TCGA-LUAD dataset with 574 samples and p biomarkers. The execution time of information searches in nine biological DB for the m-number of most relevant biomarkers with the EFS method (union of top features with five FS methods). Default parameters were used for each FS method. The 3-fold CV and 0.3 random sampling (RS) were used for model validation (V). The calculations were performed on an Intel Core i5-12400 CPU using 32 GB RAM. For notes, see text.

p	m	V	U-test	MDFS-1D	MDFS-2D	MRMR	MCFS	Ensemble	DB query
100	80	CV	00:00:05	00:00:04	00:00:04	00:00:04	00:00:13	00:00:31	00:05:44
	82	RS	00:00:02	00:00:01	00:00:02	00:00:02	00:00:05	00:00:11	00:05:26
200	140	CV	00:00:09	00:00:09	00:00:08	00:00:05	00:00:20	00:00:52	00:09:51
	149	RS	00:00:03	00:00:03	00:00:03	00:00:02	00:00:07	00:00:19	00:09:38
1000	205	CV	00:02:18	00:02:16	00:02:19	00:00:07	00:01:14	00:08:17	00:14:23
	265	RS	00:00:49	00:00:47	00:00:49	00:00:02	00:00:27	00:02:57	00:13:59

Seven metabolic and signalling pathways were selected from the KEGG and 14 pathways from the WikiPathways. Higher tissue-specific expression was observed in five genes. Twenty-six protein complex-coding genes were found. Higher tissue-specific expression was observed in 5 genes. And finally, 20 disease-related phenotypes were detected. Our analysis revealed a series of genes (SLC25A10, TGFBR1, and SFTPC, etc.) related to lung cancer.

5 Computational aspects

To test the speed efficiency of the *EnsembleFS* web app, we reviewed its performance for various data sizes. Table 1 presents example execution times of the FS and RF algorithm for previously described TCGA-LUAD RNA sequence data. The one run of the algorithm involved the following steps: calling individual or all FS algorithms, removing correlated features, estimating a ranking of the features, and calling the RF classification algorithm for random sampling (train-test split ratio of 70%–30%) or one time for the 3-fold CV method.

As shown in Table 1, the time of algorithm performance strongly depends on the FS method and hyperparameter tuning when the feature number increases. Among the applied FF, the U-test and MDFS are the fastest for the initial 100 features, while the MRMR is for 1000 features. It should be added that the execution time of the MRMR and MCFS algorithms increases if their default parameters are changed, that is, a number of relevant features for the MRMR and other cutoff methods for the MCFS. The execution time of the MDFS-2D algorithm depends on the processor's architecture (CPU or GPU).

6 Summary

We developed the *EnsembleFS* R toolkit (R package and web app) for individual FS or ensemble FS of high-dimensional molecular data and automatic collection of information on the most relevant genes from the nine well-known biological databases. In this work, we present the selected capabilities and the advantages

ICCS Camera Ready Version 2024
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-63772-8_7

of the *EnsembleFS* web application. Our results show that *EnsembleFS* is an excellent tool for selection, binary classification, and comprehensive analysis of biomarkers. We provide *EnsembleFS* as a freely accessible web server for users (30 MB limit on the total amount of data). For much larger data, we recommend using the *EnsembleFS* R package. In the current version *EnsembleFS* uses five feature filters and the random forest classifier. In the future, we plan to increase the number of feature selection and classification methods.

References

1. Breiman, L.: Random forests. Machine Learning **45**, 5–32 (2001)
2. Consortium, T.G.O.: The gene ontology resource: enriching a gold mine. Nucleic Acids Res. **49**(D1), D325–D334 (2021)
3. Determan, C.: Optimal algorithm for metabolomics classification and feature selection varies by dataset. Int. J. Biol. **7**(1) (2015)
4. Ding, C., Peng, H.: Minimum redundancy feature selection from microarray gene expression data. J. Comput. Biol. Bioinform. **3**(2), 185–205 (2005)
5. Draminski, M., Koronacki, J.: rmcfs: An r package for monte carlo feature selection and interdependency discovery. J. Stat. Softw. **85**(12), 1–28 (2018)
6. Fernandez-Delgado, M., et al.: Do we need hundreds of classifiers to solve real world classification problems? J. Mach. Learn. Technol. **15**, 3133–3181 (2014)
7. Hammerman, P., et al.: Comprehensive genomic characterization of squamous cell lung cancers. Nature **489**, 519–525 (2012)
8. He, S., et al.: MRMD3. 0: A Python tool and webserver for dimensionality reduction and data visualization via an ensemble strategy. J. Mol. Biol. p. 168116 (2023)
9. Jović, A., et al.: A review of feature selection methods with applications. 38th Int. Conv. Inf. Commun. Technol. Electron. Microelectron. (MIPRO) pp. 1200–1205 (2015)
10. Köhler, S., et al.: The Human Phenotype Ontology in 2021. Nucleic Acids Res. **49**(D1), D1207–D1217 (2021)
11. Leclercq, M., et al: Large-scale automatic feature selection for biomarker discovery in high-dimensional omics data. Front. Genet. **10**, 452 (2019)
12. Lustgarten, J., et al.: Measuring stability of feature selection in biomedical datasets. In: AMIA Annu. Symp. Proc. pp. 406–410 (2009)
13. Masoudi-Sobhanzadeh, Y., et al.: Featureselect: a software for feature selection based on machine learning approaches. BMC Bioinformatics **20**(170) (2019)
14. Mnich, K., Rudnicki, W.R.: All-relevant feature selection using multidimensional filters with exhaustive search. Inf. Sci. **524**, 277–297 (2020)
15. Okuda, S., et al.: Kegg atlas mapping for global analysis of metabolic pathways. Nucleic Acids Res. **36**, W423–6 (2008)
16. Polewko-Klim, A., Rudnicki, W.R.: Analysis of ensemble feature selection for correlated high-dimensional rna-seq cancer data. In: Computational Science-ICCS 2020: 20th Int. Conf, Amsterdam, The Netherlands, June 3-5, 2020, Proceedings, Part III 20. pp. 525–538. Springer (2020)
17. Rohart, F., et al.: mixOmics: An R package for 'omics feature selection and multiple data integration. PLoS Comput. Biol. **13**(11), e1005752 (2017)
18. Stawiski, K., et al.: OmicSelector: automatic feature selection and deep learning modeling for omic experiments. bioRxiv (2022). `https://doi.org/10.1101/2022.06.01.494299`

ICCS Camera Ready Version 2024
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-63772-8_7