Analysis of Ensemble Feature Selection for Correlated High-Dimensional RNA-Seq Cancer Data

Aneta Polewko-Klim[1]([✉]) [ID] and Witold R. Rudnicki[1,2,3] [ID]

[1] Institute of Informatics, University of Białystok, Białystok, Poland
anetapol@uwb.edu.pl
[2] Computational Center, University of Białystok, Białystok, Poland
[3] Interdisciplinary Centre for Mathematical and Computational Modelling,
University of Warsaw, Warsaw, Poland

Abstract. Discovery of diagnostic and prognostic molecular markers is important and actively pursued the research field in cancer research. For complex diseases, this process is often performed using Machine Learning. The current study compares two approaches for the discovery of relevant variables: by application of a single feature selection algorithm, versus by an ensemble of diverse algorithms. These approaches are used to identify variables that are relevant discerning of four cancer types using RNA-seq profiles from the Cancer Genome Atlas. The comparison is carried out in two directions: evaluating the predictive performance of models and monitoring the stability of selected variables. The most informative features are identified using a four feature selection algorithms, namely U-test, ReliefF, and two variants of the MDFS algorithm. Discerning normal and tumor tissues is performed using the Random Forest algorithm. The highest stability of the feature set was obtained when U-test was used. Unfortunately, models built on feature sets obtained from the ensemble of feature selection algorithms were no better than for models developed on feature sets obtained from individual algorithms. On the other hand, the feature selectors leading to the best classification results varied between data sets.

Keywords: Random forest · RNA · Feature selection · Ensemble learning

1 Introduction

The high-throughput DNA sequencing techniques produce data with tens of thousands probes and each of them could be potentially relevant for diagnostics, prognosis and therapeutics.

Feature selection (FS) techniques are indispensable tools for filtering out irrelevant variables and ranking the relevant ones in molecular biological investigations [15,28]. The choice of the FS method is very important for further

investigation because it greatly limits number of features under scrutiny, allowing to concentrate on most relevant ones. On the other hand, FS increases the risk of omitting biological important variables.

FS methods are typically divided into three major groups, namely filters, wrappers, and embedded [1]. The bias in the filtering FS methods does not correlate with the classification algorithms, hence they generalise better than the other methods. Nevertheless, it is well known that individual feature selection algorithms are not robust with respect to fluctuations in the input data [22]. Consequently, application of a single FS algorithm cannot ensure optimal modelling results both in terms of predictive performance and stability. This is particularly evident in the integrative analysis of high-dimensional *omics data [18].

There are numerous FS algorithms that are based on different principles and can generate highly variable results for the same data set. The presence of highly correlated features may result in multiple equally optimal set of features and consequently to the instability of FS method [10]. Such instability reduces the confidence in selected features [22] and their usage as diagnostic or prognostic markers. This variability can be to some extent minimised by application of ensemble methods (EFS) that involve combination of different selectors [1].

1.1   Related Work

The ensemble FS can be broadly assigned to one of two classes: homogeneous (the same base feature selector) and heterogeneous (different feature selectors), [1]. Regardless of the class, the output of ensemble FS is given either in a form of a final feature set or as a ranking of features. Therefore some papers focus on the comparison of different strategies for the ordering of these feature subsets [29]. Other researchers are focused on the evaluation of ensembles. Two quantities of interest are the diversity [25] and stability of the feature selection process [20,22]. And though various methods of feature selection have been developed for high-dimensional data, such as high-throughput genomics data, it is still a big challenge to choose the appropriate method for this type of data [15,30].

The stability of FS algorithms for the classification of this type of data has been investigated for instance by Moulos et al. [20] and Dessi and Pes [6]. It was shown, that stability of ensemble feature selection increase only for these FS methods that are intrinsically weak (in term of stability). Shahrjooihaghighi et al. [26] proposed an ensemble FS based on the fusion of five feature selection methods (rank product, fold change ratio, ABCR, t-test, and PLSDA) for more effective biomarker discovery. The methodology for comparing the outcomes of different FS techniques is presented in [5].

Current study is focused on developing and optimisation of a feature selection protocol aiming at identification of biomarkers important for diagnostic of cancers using the results of high-throughput molecular biology experimental methods. It is based on ensemble of four diverse feature selection methods and application of classification algorithm that is used to evaluate quality of the set of features.

The protocol was applied to analyse four human cancer tumor types from The Cancer Genome Atlas (TCGA, https://www.cancer.gov/tcga).

In particular the following detailed issues were explored:

– whether application of ensemble of FS methods gives more stable results than individual algorithms;
– what is the optimal number of variables for individual algorithms and for ensemble;
– whether models built using features returned by ensemble are better than models built using the same number; of variables returned by individual algorithms;
– which feature selection algorithm returns best sets of variables?

The main contributions of the current study are following:

– we present a novel perspective of optimization and evaluation of the feature selection for correlated high-dimensional RNA-Seq cancer data;
– we compare both the predictive performance of models and the stability of selected feature sets in ensemble feature selection with that of individual FS algorithms;
– we show, that performance of feature selection methods vary between data sets even in for very similar data sets;
– we propose to use an ensemble approach as a reference for selecting the method that works best for a particular data set.

2 Materials and Methods

2.1 Data

Four data sets from The Cancer Genome Atlas database that contain RNA-sequencing data of tumor-adjacent normal tissues for various typed of cancer were used. [3,4,8,12,14,31] These data set all include a large number of highly correlated and potentially informative features [21]. The preprocesing of data involved standard steps for RNA-Seq data. First the log2 transformation was performed. Then features with zero and near zero (1%) variance across patients were removed. After preprocessing the datasets contain:

– the primary BRCA dataset: 1205 samples (112 normal and 1093 tumor), 20223 variables;
– the LUAD dataset: 574 samples (59 normal and 515 tumor), 20172 variables;
– the KIRC dataset: 605 samples (72 normal and 533 tumor), 20222 variables,
– the HNSC dataset: 564 samples (44 normal and 520 tumor), 20235 variables.

All data sets are imbalanced, they contain roughly ten times more cancer than normal samples.

2.2   Methods

Filters Used for Feature Selection. The procedure outlined above was applied to four filter FS methods, namely, Mann-Whitney U-test [17], ReliefF [11,13] and MDFS [19,23] in two variants: one-dimensional (MDFS-1D) and two-dimensional (MDFS-2D). Since only the ranking of variables is used in the procedure outlined above no corrections of p-value due to multiple testing were necessary.

U-test is a robust statistical filter that is routinely used in analysis of *omics data. It assigns probability to the hypothesis that two samples corresponding to two decision classes (normal and tumor tissue) are drawn from populations with the same average value. The U-test use the p-value to select and rank the features.

MDFS is a filter, which is based on information theoretical approach, and which can take synergistic effects between variables into account [19,23]. MDFS also uses p-values of the test to rank features. In the current study, 1D and 2D version of MDFS algorithm were used, referred to as MDFS-1D and MDFS-2D, respectively.

ReliefF is a filter that computes ranking of importance for variables in the information system, based on the distances in the small-dimensional subspaces of the system [13]. Two variants of distance between nearest neighbours, namely, *ReliefFexpRank* and *ReliefFbestK* were tested for the current study. Slightly better results were obtained for the former, hence it was used in all subsequent work. This R implementation of algorithm from *CORElearn* package was used [24].

Filter-Based Feature Selection. The individual prediction models in k-fold cross-validation for each of four filter FS methods and data sets were constructed. The feature selection process and the learning process from RNA-Seq data set were realized by using the Algorithm 1.

This algorithm outlined above was repeated for several values of N and it was repeated multiple times, to minimize the effects of random fluctuations. The stability of feature selection was measured by comparing feature sets obtained in multiple runs of the procedure.

Ensemble Feature Selection. The ensemble set of N-top relevant variables was constructed by a union of top-N variables from each filter FS methods, as it is shown in Algorithm 2. The size of the set may vary between N and $4N$ depending on the similarity of the sets returned by individual FS algorithms. All comparisons between feature sets obtained from ensemble and features sets obtained by individual filters were performed on sets with comparable numbers of total variables. For example if union of four top-5 sets resulted in a set with 20 variables it was subsequently compared with other sets containing 20 variables.

Algorithm 1: FS$(l, f, N, D = \{S_1, \ldots, S_k\})$ the filter feature selection algorithm with Random forest classifier

---

    input : Learning method l
               Feature selection method f
               Number of top features N
               Data set $D = \{(y_i, x_i)\}_{i=1}^{M}$ with $V = \{v_1, \ldots, v_p\}$ features
               and with M instances, randomly partitioned into about
               equally-sized folds S_j
1 **output:** Ranked feature sets F_j, $j = 1, \ldots, k$
               Performance estimation metric E
2 **foreach** S_j **do**
3   Define the training set $D_{\setminus j}(V) \leftarrow D(V) \setminus S_j(V)$
4   Perform feature selection on the training set $R_j \leftarrow f(D_{\setminus j}(V))$
5   Remove highly correlated features with ranked list R_j of features
6   Collect the N highest ranked feature set $F_j = \{v_1, \ldots, v_n\}$ with R_j
7   Build the model on the training set $L_j \leftarrow l(D_{\setminus j}(F_j))$
8   Performance estimation: use the trained model L_j on a test set S_j
9 $E \leftarrow \frac{1}{k} \Sigma E_j$
10 **end**

Algorithm 2: EFS$(l, W = \{F_{11}, \ldots, F_{4k}\}, D = \{S_1, \ldots, S_k\})$ the ensemble feature selection algorithm with Random forest classifier

---

    input : Learning method l
               The $4 \times$ k sets of top-N uncorrelated features with 4 filters W_i
               $F_{i,j}, i = 1, \ldots, 4, j = 1, \ldots, k$ with top-N features
               Data set $D = \{(y_m, x_m)\}_{m=1}^{M}$ described with features $F_{i,j}$
               and with M instances, randomly partitioned into about
               equally-sized folds S_n
1 **output:** Collected feature sets C_p, $p = 1, \ldots, k$
               Performance estimation metric E_p
2 **foreach** S_j **do**
3   Collect the union of feature set $C_p = F_{1j} \cup F_{2j} \cup F_{3j} \cup F_{4j}$
4   Define the training set $D_{\setminus n}(C_p) \leftarrow D(C_p) \setminus S_n(C_p)$
5   Build the model on the training set $L_{n,p} \leftarrow l(D_{\setminus n}(C_p))$
6   Performance estimation: $E_{n,p} \leftarrow L_{n,p}(S_n(C_p))$
7 $E_p \leftarrow \frac{1}{k} \Sigma E_{n,p}$
8 **end**

All applied filters provide their own ranking of the features. The U-test and MDFS algorithms rank features by their statistical significance and, the ReliefF algorithm by their performance in classification. Then the joint set of most important variables is created as a union of top-N sets from individual rankings. The ranking within the combined set was not necessary and it was never performed.

Algorithm 2 was repeated several times for different values of N as in the case Algorithm 1. The stability of feature selection was also estimated.

Classification. The quality of the feature-set was evaluated by building a machine learning model using selected features and measuring its quality. To this end the Random Forest [2] algorithm was used. It has been shown that Random Forest is generally reliable algorithm, that works well out-of the box, rarely fails and in most cases returns results that are very close to best achievable for given problem [7]. The quality of model was evaluated using area under ROC curve (AUC). This measure is independent of the balance of classes in the data and does not need any fitting. The scheme of ensemble feature selection and supervised classification is presented in Fig. 1.
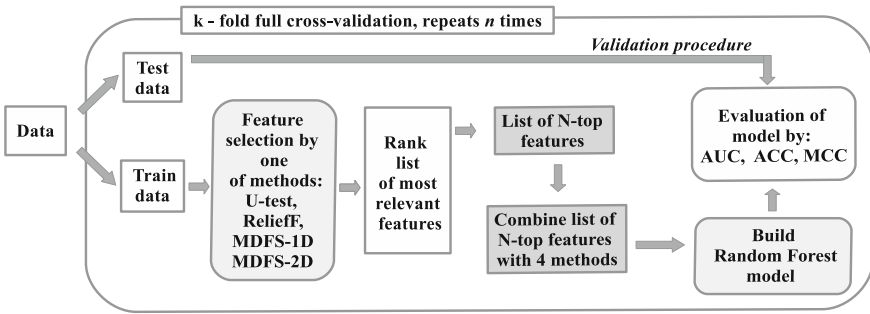

Fig. 1. Pipeline of the ensemble FS method. See notation in text.

Measuring Stability of Feature Selection. The total stability of filter FS method is measured as the average of the pairwise similarity for all pairs of the most informative feature subsets (s_i, s_j) from n runs of a model in full k-fold cross-validation. To this end the Lustgarten's stability measure (ASM) [16], which can be applied to sets of unequal sizes, was used. It is described by the formula:

$$ASM = \frac{2}{c(c-1)} \sum_{i=1}^{c-1} \sum_{j=i+1}^{c} \left(\frac{|s_i \cap s_j| - |s_i| * |s_j| / m}{min(|s_i|, |s_j|) - max(0, |s_i| + |s_j| - m)} \right) \quad (1)$$

where: m is total feature number of dataset and $c = n*k$.

Optimization of Feature Selection. In the first step four threshold levels for defining highly correlated variables were examined to establish threshold leading to best results of classification. Four thresholds levels were tested: $|r| = \{0.7, 0.75, 0.8, 0.9\}$. The subsequent analyses were performed for the optimal threshold level.

The following analyses were performed for each individual FS filter and for ensemble FS filters:

– how many uncorrelated variables should be included in the model to obtain best classification;
– how stable is stability measure for top-N feature subsets;
– whether adding the highly correlated variables to top-N variables influences predictive power.

Entire modelling protocol, including bot feature selection and model building step was performed within $k = 5$ fold cross-validation and was repeated $n = 30$ times, independently for each FS method and data set. Within each cross-validation iteration feature selection algorithm was performed once and then models were trained for all feature set sizes $N = \{5, 10, \ldots, 200\}$.

Analysis was performed using the R (version 3.5) [27] and R/Bioconductor packages [9].

3 Results and Discussion

3.1 Model Accuracy

In the first step, the impact of correlation between informative features on the predictive power of RF model was examined. The results of this analysis are displayed in Fig. 2. It can be seen that squares corresponding to correlation threshold 0.7 in many cases fall bellow other lines on the AUC plots. Therefore threshold for removal of highly correlated variables was set at Spearman's correlation coeffcient r higher than 0.75. This value of coefficient is applied in the subsequent analysis. One may note, that MDFS-1D filter is the most robust with respect to change in the feature level correlation among the applied FS methods.

At the next stage of the analysis, the accuracy of models built using top-N features was examined, see Fig. 3. The number of variables for ensemble model is obtained as the average number of variables in the union of top-N variables from all FS methods averaged over 150 cross-validated sets. Generally, the performance of the models is poor for the smallest sizes of variable sets but increases rapidly with increasing number of variables, reaching plateau after roughly 40 variables are included. However, there are notable exceptions. In particular for BRCA the best model was obtained with 20 variables returned by U-test. Even stronger effect was obtained for KIRC data set. Here the MDFS-2D feature selection leads to clearly best results at 25 variables, whereas the best results for other filters are obtained with 15 variables.
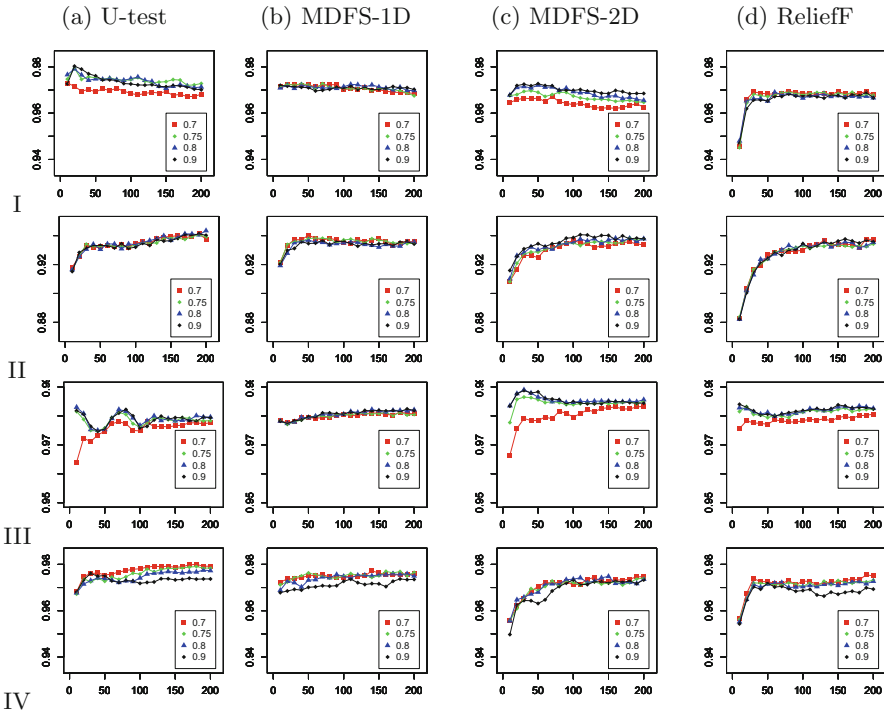
Fig. 2. AUC vs N-top biomarkers for different value of Spearman's rank correlation coefficient. Results for BRCA, HNSC, KIRC and LUAD data sets are displayed in rows I, II, III, and IV, respectively.

Relative performance of models developed using different FS algorithms vary significantly between data sets. For example, the MDFS-2D is clearly the best feature selector for KIRC, and that strongly suggest that non-trivial synergies between variables are present in this data set. MDFS-1D is best feature selector for HNSC data set, U-test is best for BRCA, both algorithms are similarly good for LUAD.

In all cases, the AUC values of models built using variables returned by an ensemble of FS algorithms are comparable to individual models built with a similar number of variables, the AUC curves of the ensemble models (full circle points) are located roughly in the middle of other models as shown in Fig. 3.

Only for BRCA data set for the number of variables larger than 40 the performance of model built on ensemble variables is comparable with the best model for the individual data set, which in this case is a model built using variables from U-test.
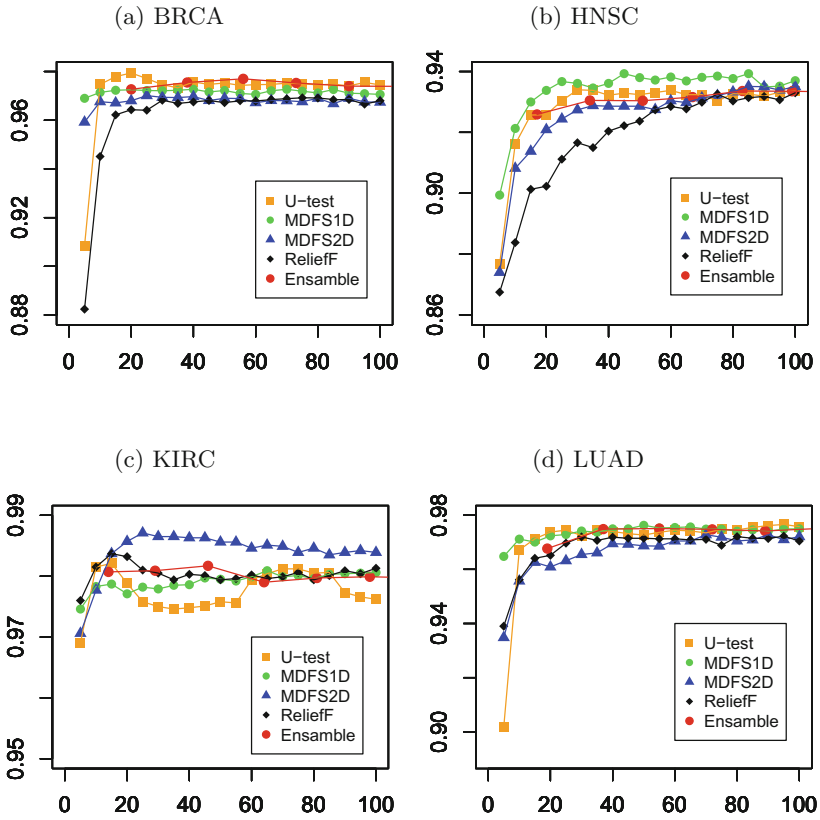
Fig. 3. AUC for models built using top-N variables.

In the next step the effect of adding back redundant variables was examined. To this end the RF models were built for the sets of variables consisting of uncorrelated top-N variables and all informative variables highly correlated with them that were previously removed from feature rank list. The results are displayed in Fig. 4. Clearly adding redundant variables to the main feature set does not improve classification results in most cases. An exception are models built using variables obtained with the MDFS-2D method for BRCA, HNSC and KIRC data. This effect may arise due to inclusion of correlated variables which interact synergistically with other variables in a slightly different way than those previously included.
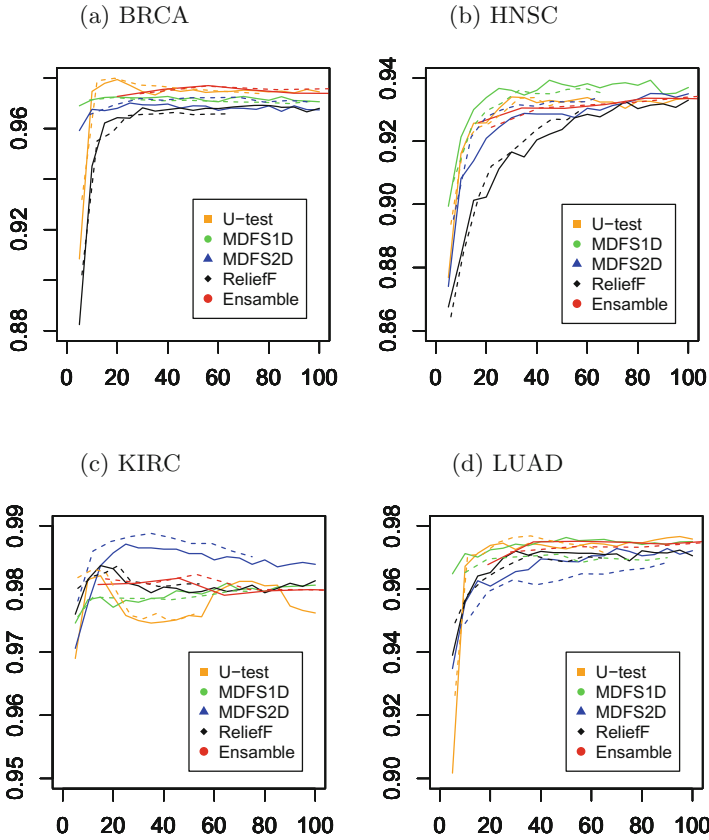
Fig. 4. AUC for models built using top-N variables. Solid lines correspond to models built using top-N uncorrelated variables. Dashed lines correspond to models built using top-N uncorrelated variables and variables correlated with them.

3.2   Stability of Variable Sets

Often the important property of a feature selection method is stability or robustness of the selected features to perturbations in the data. This is particularly important for identification of prognostic or diagnostic markers. Therefore the sensitivity of feature selection algorithms to variations in the training sets that arise in the cross-validation were examined. The similarity between 150 feature subsets obtained in 150 iterations of cross-validation were measured using the Lustgarten's index ASM, see Fig. 5. The highest stability was obtained for variables selected with U-test. For this FS method the value of the ASM index varies between 0.7 and 0.8. The remaining FS methods are much less stable, with least stable MDFS-2D for which the ASM index is generally below 0.2 and most stable ReliefF for which ASM index varies between 0.3 and 0.5. The difference in stability between algorithms is due to the differences in approach used
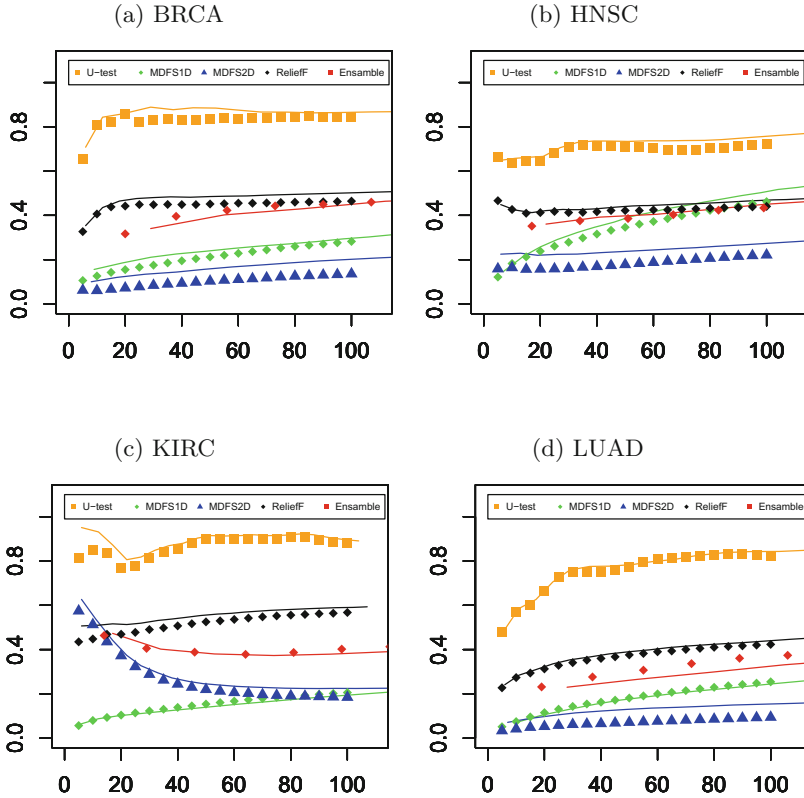
(a) BRCA   (b) HNSC

(c) KIRC   (d) LUAD

Fig. 5. Clock-wise the average similarity (ASM) between 150 feature subsets for top-N variables. Dotted lines correspond to sets consisting from top-N variables. Solid lines correspond to sets consisting from top-N variables and variables highly correlated with them.

by the algorithm. The U-test is a deterministic algorithm, for which differences arise exclusively due to variation of sample composition. On the other hand all other remaining algorithms rely on randomisation, hence increased variance can be expected. In most cases the stability increases with increasing number of variables. The notable exception is the MDFS-2D algorithm for KIRC data set, where relatively high stability (ASM > 0.5) is achieved for smallest size of feature set, and then it rapidly drops with increasing number of variables. This result is obtained for the same data set, where models built on feature sets returned by MDFS-2D have highest predictive quality. This result strongly suggest existence of a small core of most relevant variables, that must be present in nearly all cases and that strongly contribute to classification. This small core is augmented by a diverse group of loosely correlated relevant but redundant variables. Finally, in most cases adding redundant variables increases the stability of feature subsets, but the difference is small.

3.3   Computational Aspects

The training time does not depend on the type of molecular data, such as microarray gene expression data or DNA copy number data. Execution time of the task depends on the size of the dataset, the number of training iterations, the feature selection algorithm, as well as the CPU model or the GPU model. The most time consuming individual step of the algorithm is feature selection, the model building with Random Forest is relatively quick. However, 150 distinct Random Forest were built using the results of the same feature selection step. Therefore total time of both components was similar. The example execution times for a single iteration of algorithm for KIRC data set are presented in the Table 1. Among feature selection algorithms used in the study, the ReliefF is by far the most time-consuming.

Table 1. Execution times for a single iteration of the algorithm for the KIRC data set. Computations performed on a CPU Intel Xeon Processor E5-2650v2. The MDFS-2D algorithm was executed using a GPU-accelerated version on NVIDIA Tesla K80 co-processor.

U-test	MDFS-1D	MDFS-2D	ReliefF	Ensemble	RF ×1	RF × 100	Total
00 m:37 s	00 m:04 s	00 m:03 s	05 m:41 s	05 m:54 s	00 m:03 s	05 m:19 s	10 m:13 s

The single run of the algorithm involved calling four FS methods, removing correlated features, producing a ranking of the features, and calling RF classification algorithm 150 times (5 feature sets × 20 sizes of feature set). The algorithm was executed 150 times, computations for one data set took about 25 h of CPU time.

4   Conclusions

The current study demonstrates that relying on a single FS algorithm is not optimal. Different FS algorithms are best suited for identification of the most relevant feature sets in various data sets. Combining variables from multiple FS algorithms into a single feature set does not improve the performance in comparison with an equally numerous feature set generated by the individual algorithm that is best suited for the particular data set. On the other hand, application of multiple algorithms increases the chances of identifying the best FS algorithm for the problem under scrutiny. In particular, application of the FS algorithms that can detect synergies in the data can significantly improve the quality of machine learning models.

Interestingly, the stability of a FS algorithm is not required for building a highly predictive machine learning models. This is possible, since biological systems often contain multiple informative variables. Therefore, useful models can be obtained using very diverse combinations of predictive variables.

Notes

References

1. Bolón-Canedo, V., Alonso-Betanzos, A.: Ensembles for feature selection: a review and future trends. Inf. Fus. **52**, 1–12 (2019)
2. Breiman, L.: Random forests. Mach. Learn. **45**(5), 5–32 (2001)
3. Ciriello, G., Gatza, M.L., Beck, A.H., Wilkerson, M.D., et al.: Comprehensive molecular portraits of invasive lobular breast cancer. Cell **163**(2), P506–P519 (2015)
4. Collisson, E., Campbell, J., Brooks, A., et al.: Comprehensive molecular profiling of lung adenocarcinoma. Nature **511**, 543–550 (2014)
5. Dessí, N., Pascariello, E., Pes, B.: A comparative analysis of biomarker selection techniques. BioMed Res. Int. **2013**, 387673 (2013)
6. Dessì, N., Pes, B.: Stability in biomarker discovery: does ensemble feature selection really help? In: Ali, M., Kwon, Y.S., Lee, C.-H., Kim, J., Kim, Y. (eds.) IEA/AIE 2015. LNCS (LNAI), vol. 9101, pp. 191–200. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-19066-2_19
7. Fernández-Delgado, M., Cernadas, E., Barro, S., Amorim, D.: Do we need hundreds of classifiers to solve real world classification problems? J. Mach. Learn. Res. **15**, 3133–3181 (2014)
8. Hammerman, P., Lawrence, M., Voet, D., et al.: Comprehensive genomic characterization of squamous cell lung cancers. Nature **489**, 519–525 (2012)
9. Huber, W., Carey, V., Gentleman, R., Anders, S., et al.: Orchestrating high-throughput genomic analysis with Bioconductor. Nat. Methods **12**(2), 115–121 (2015)
10. Kamkar, I., Gupta, S.K., Phung, D., Venkatesh, S.: Exploiting feature relationships towards stable feature selection. In: 2015 IEEE International Conference on Data Science and Advanced Analytics (DSAA), pp. 1–10 (2015)
11. Kira, K., Rendell, L.: The feature selection problem: Traditional methods and a new algorithm. AAAI **2**, 129–134 (1992)
12. Koboldt, D., Fulton, R., McLellan, M., et al.: Comprehensive molecular portraits of human breast tumours. Nature **490**, 61–70 (2014)
13. Kononenko, I.: Estimating attributes: analysis and extensions of RELIEF. In: Bergadano, F., De Raedt, L. (eds.) ECML 1994. LNCS, vol. 784, pp. 171–182. Springer, Heidelberg (1994). https://doi.org/10.1007/3-540-57868-4_57
14. Lawrence, M., Sougnez, C., Lichtenstein, L., et al.: Comprehensive genomic characterization of head and neck squamous cell carcinomas. Nature **517**, 576–582 (2015)
15. Liang, S., Ma, A., Yang, S., Wang, Y., Ma, Q.: A review of matched-pairs feature selection methods for gene expression data analysis. Comput. Struct. Biotechnol. J. **16**, 88–97 (2018)
16. Lustgarten, J.L., Gopalakrishnan, V., Visweswaran, S.: Measuring stability of feature selection in biomedical datasets. In: AMIA Annual Symposium Proceedings, pp. 406–410. AMIA (2009)
17. Mann, H., Whitney, D.: Controlling the false discovery rate: a practical and powerful approach to multiple testing. Ann. Math. Stat. **18**(1), 50–60 (1947)

18. Meng, C., Zeleznik, O.A., Thallinger, G.G., et al.: Dimension reduction techniques for the integrative analysis of multi-omics data. Briefings Bioinf. **17**(4), 628–641 (2016)
19. Mnich, K., Rudnicki, W.: All-relevant feature selection using multidimensional filters with exhaustive search. Inf. Sci. **524**, 277–297 (2020)
20. Moulos, P., Kanaris, I., Bontempi, G.: Stability of feature selection algorithms for classification in high-throughput genomics datasets. In: 13th IEEE International Conference on BioInformatics and BioEngineering, pp. 1–4 (2013)
21. Peng, L., Bian, X.W., Li, D.K., et al.: Large-scale RNA-Seq transcriptome analysis of 4043 cancers and 548 normal tissue controls across 12 TCGA cancer types. Sci. Rep. **5**(1), 1–18 (2015)
22. Pes, B.: Ensemble feature selection for high-dimensional data: a stability analysis across multiple domains. Neural Comput. Appl. **32**(10), 5951–5973 (2019). https://doi.org/10.1007/s00521-019-04082-3
23. Piliszek, R., Mnich, K., Migacz, S., et al.: MDFS: multidimensional feature selection in R. R J. **11**(1), 197–210 (2019)
24. Robnik-Sikonja, M., Savicky, P.: CORElearn: classification, regression and feature evaluation, R package version 1.54.1 (2018). https://CRAN.R-project.org/package=CORElearn
25. Seijo-Pardo, B., Bolón-Canedo, V., Alonso-Betanzos, A.: Testing different ensemble configurations for feature selection. Neural Process. Lett. **46**(3), 857–880 (2017)
26. Shahrjooihaghighi, A., Frigui, H., Zhang, X., Wei, X., Shi, B., Trabelsi, A.: An ensemble feature selection method for biomarker discovery. In: 2017 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT), pp. 416–421 (2017)
27. Team, R.C.: R: A language and environment for statistical computing. R Foundation for Statistical Computing (2017). https://www.R-project.org/
28. Vanjimalar, S., Ramyachitra, D., Manikandan, P.: A review on feature selection techniques for gene expression data. In: 2018 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC), pp. 1–4. IEEE, ICCIC (2018)
29. Wang, J., Xu, J., Zhao, C., Peng, Y., Wang, H.: An ensemble feature selection method for high-dimensional data based on sort aggregation. Syst. Sci. Control Eng. **7**(2), 32–39 (2019)
30. Wenric, S., Shemirani, R.: Using supervised learning methods for gene selection in RNA-seq case-control studies. Front. Genet. **34**(4), 301–312 (2018)
31. Zhou, Y., Zhou, B., Pache, L., Chang, M., et al.: Metascape provides abiologist-oriented resource for the analysis of systems-level datasets. Nat. Commun. **10**(1), 1–10 (2019)