EFSIS: Ensemble Feature Selection Integrating Stability

Xiaokang Zhang, Inge Jonassen

Computational Biology Unit, Department of Informatics, University of Bergen, Norway

Preprint paper

Introduction

Ensemble learning has been widely applied in pattern recognition with the benefit of more robust and accurate than the individual learners. This ensemble logic has recently also been more and more applied in feature selection. There are basically two strategies for ensemble feature selection, namely data perturbation and function perturbation. Data perturbation subsamples the whole dataset, and then selects the features consistently ranked highly across those data subsets to improve both the stability of the selector and the prediction accuracy for a classifier. Function perturbation integrates multiple selectors to free the user from having to decide on the most appropriate selector for any given situation, while maintaining or improving classification performance. We here propose a framework, EFSIS, combining these two strategies. Empirical results indicate that EFSIS gives both high prediction accuracy and stability.

Methods

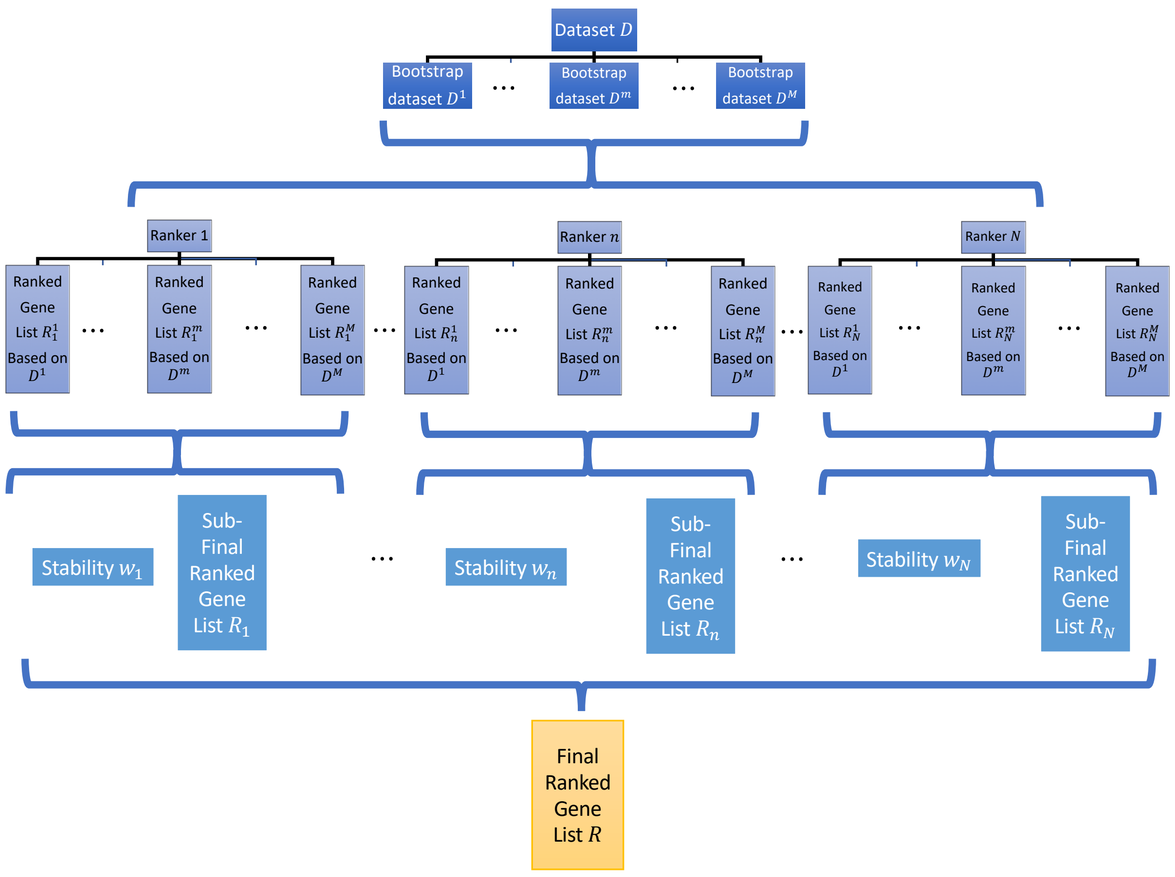

Figure 1. Work scheme of our method «EFSIS»

Aggregation strategy

There are two aggregations in the paradigm as Fig. 1 shows.

- To aggregate the rankings of one selection method from different sub samplings, we use the product of the ranks of one feature across the feature selection (ranked) lists as its aggregated ranking score [1].
- To aggregate the results from different methods (SAM, GeoDE, ReliefF and Information Gain [2 - 5]), we apply the product strategy with the stability as a weight.

Datasets

We tested our method on 6 different cancer-related transcriptomics (see Table 1). These datasets represent a broad range of characteristics in terms of biological information (different types of cancers), number of samples and number of attributes (genes).

Dataset name	Attributes	Samples
AML [6]	12625	54
CNS [7]	7129	60
ColonBreast [8]	22283	52
DLBCL [9]	7129	77
Leukemia [10]	7129	72
ProstateSingh [11]	12600	102

Table 1. Datasets

Evaluation

We used a 10-fold cross-validation ($K = 10$ in Fig. 2) scheme for all the feature selection methods by applying them to nine tenths of the samples (training set) and assessing the prediction accuracy on the remaining (test set). The model performance was then quantified using the Area Under Curve (AUC) measure, and summarized over the 10 runs performed in the cross-validation.

Tests were performed with varying number of features. For classification we used Support Vector Machine (with a linear kernel).

We compared our method EFSIS with the integrated individual methods and function perturbation.

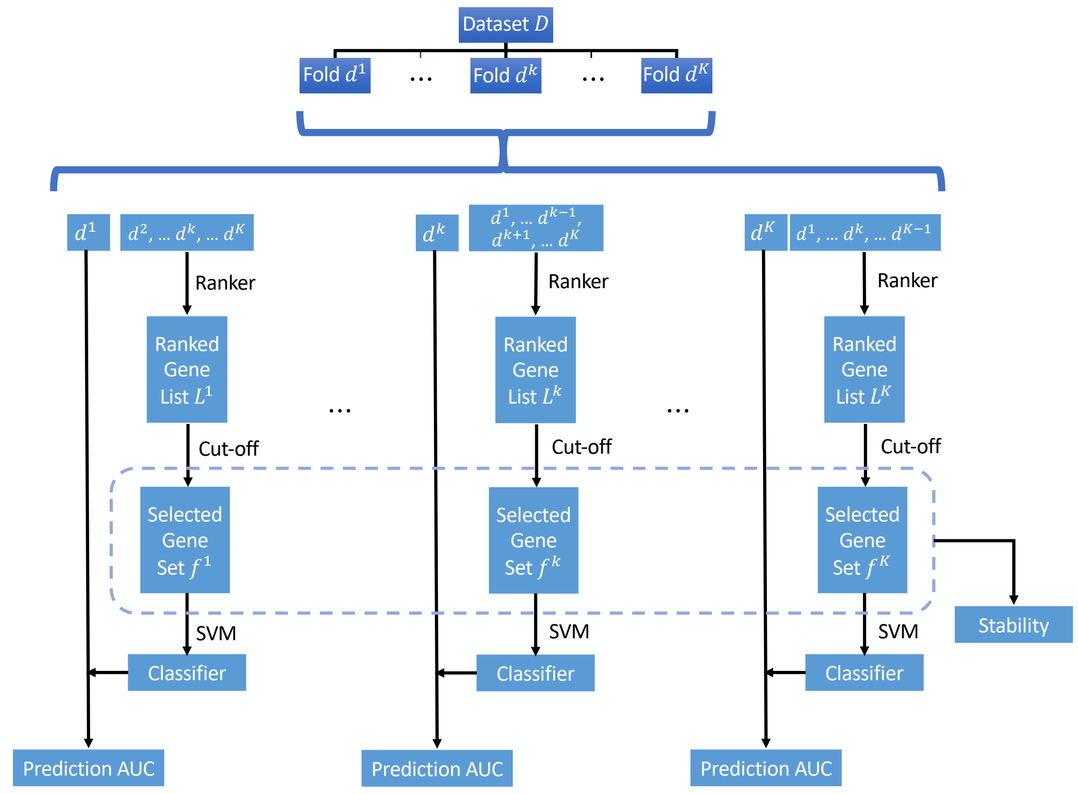

Figure 2. Workflow of performance evaluation

Results

Dataset	Ranker	Percentage of selected features (%) 0.3	0.5	0.7	1	1.5	2	3	4	5
AML	SAM	0.69 ± 0.17	0.73 ± 0.16	0.73 ± 0.20	0.76 ± 0.14	0.78 ± 0.17	0.75 ± 0.18	0.74 ± 0.20	0.76 ± 0.17	0.77 ± 0.16
	GeoDE	0.74 ± 0.16	0.69 ± 0.25	0.76 ± 0.20	0.76 ± 0.16	0.80 ± 0.16	0.80 ± 0.18	0.84 ± 0.15	0.79 ± 0.18	0.79 ± 0.22
	ReliefF	0.66 ± 0.14	0.67 ± 0.16	0.66 ± 0.21	0.66 ± 0.21	0.70 ± 0.17	0.74 ± 0.15	0.71 ± 0.17	0.71 ± 0.17	0.71 ± 0.17
	Info_Gain	0.81 ± 0.16	0.75 ± 0.18	0.74 ± 0.16	0.73 ± 0.17	0.79 ± 0.14	0.76 ± 0.17	0.77 ± 0.17	0.79 ± 0.16	0.80 ± 0.17
	Func_Pert	0.74 ± 0.20	0.74 ± 0.17	0.74 ± 0.16	0.77 ± 0.16	0.78 ± 0.15	0.75 ± 0.16	0.79 ± 0.20	0.75 ± 0.19	0.78 ± 0.13
	EFSIS	0.71 ± 0.16	0.70 ± 0.19	0.74 ± 0.16	0.78 ± 0.12	0.73 ± 0.17	0.76 ± 0.17	0.79 ± 0.14	0.76 ± 0.17	0.76 ± 0.17
CNS	SAM	0.72 ± 0.22	0.69 ± 0.22	0.71 ± 0.20	0.72 ± 0.17	0.71 ± 0.21	0.72 ± 0.17	0.73 ± 0.18	0.69 ± 0.19	0.73 ± 0.18
	GeoDE	0.63 ± 0.16	0.76 ± 0.08	0.81 ± 0.16	0.82 ± 0.13	0.82 ± 0.18	0.88 ± 0.17	0.89 ± 0.16	0.88 ± 0.14	0.90 ± 0.14
	ReliefF	0.72 ± 0.15	0.74 ± 0.17	0.78 ± 0.20	0.78 ± 0.16	0.76 ± 0.17	0.75 ± 0.18	0.69 ± 0.16	0.70 ± 0.16	0.70 ± 0.18
	Info_Gain	0.69 ± 0.17	0.78 ± 0.18	0.76 ± 0.19	0.71 ± 0.19	0.65 ± 0.17	0.70 ± 0.13	0.66 ± 0.18	0.71 ± 0.16	0.78 ± 0.15
	Func_Pert	0.75 ± 0.11	0.70 ± 0.17	0.70 ± 0.16	0.76 ± 0.18	0.78 ± 0.16	0.74 ± 0.21	0.77 ± 0.20	0.81 ± 0.21	0.78 ± 0.15
	EFSIS	0.67 ± 0.14	0.66 ± 0.18	0.73 ± 0.14	0.75 ± 0.17	0.80 ± 0.15	0.72 ± 0.21	0.72 ± 0.19	0.74 ± 0.19	0.75 ± 0.18
ColonBreast	SAM	0.98 ± 0.08	0.98 ± 0.08	0.97 ± 0.06	0.98 ± 0.05	0.99 ± 0.03	0.97 ± 0.07	0.97 ± 0.06	0.97 ± 0.06	0.97 ± 0.06
	GeoDE	0.99 ± 0.04	0.99 ± 0.04	0.99 ± 0.04	0.98 ± 0.05	0.95 ± 0.08	0.95 ± 0.08	0.95 ± 0.08	0.95 ± 0.08	0.98 ± 0.05
	ReliefF	0.92 ± 0.14	0.92 ± 0.12	0.93 ± 0.11	0.94 ± 0.12	0.94 ± 0.12	0.94 ± 0.12	0.95 ± 0.09	0.95 ± 0.08	0.95 ± 0.08
	Info_Gain	0.98 ± 0.05	0.95 ± 0.08	0.95 ± 0.08	0.95 ± 0.08	0.98 ± 0.08	0.98 ± 0.08	0.99 ± 0.04	0.98 ± 0.08	0.95 ± 0.12
	Func_Pert	0.98 ± 0.08	0.99 ± 0.04	0.98 ± 0.05	0.97 ± 0.05	0.97 ± 0.06	0.99 ± 0.04	0.98 ± 0.05	0.99 ± 0.03	1.00 ± 0.00
	EFSIS	0.96 ± 0.09	0.97 ± 0.08	0.98 ± 0.08	0.99 ± 0.04	0.98 ± 0.05	0.98 ± 0.05	0.98 ± 0.05	0.98 ± 0.05	0.99 ± 0.04
DLBCL	SAM	0.91 ± 0.13	0.90 ± 0.12	0.96 ± 0.08	0.96 ± 0.06	0.97 ± 0.07	0.97 ± 0.07	0.95 ± 0.11	0.94 ± 0.11	0.97 ± 0.07
	GeoDE	0.86 ± 0.10	0.87 ± 0.10	0.86 ± 0.12	0.89 ± 0.10	0.88 ± 0.16	0.86 ± 0.22	0.88 ± 0.14	0.89 ± 0.11	0.92 ± 0.10
	ReliefF	0.91 ± 0.15	0.95 ± 0.10	0.96 ± 0.08	0.97 ± 0.06	0.98 ± 0.04	0.99 ± 0.03	0.98 ± 0.04	0.99 ± 0.03	0.99 ± 0.03
	Info_Gain	0.95 ± 0.11	0.95 ± 0.09	0.95 ± 0.11	0.96 ± 0.08	0.96 ± 0.08	0.96 ± 0.08	0.96 ± 0.08	0.97 ± 0.08	0.96 ± 0.06
	Func_Pert	0.91 ± 0.10	0.93 ± 0.09	0.94 ± 0.08	0.96 ± 0.07	0.96 ± 0.07	0.96 ± 0.07	0.95 ± 0.11	0.98 ± 0.06	0.95 ± 0.11
	EFSIS	0.93 ± 0.10	0.93 ± 0.10	0.96 ± 0.08	0.94 ± 0.09	0.97 ± 0.06	0.98 ± 0.05	0.98 ± 0.05	0.98 ± 0.05	0.98 ± 0.05
Leukemia	SAM	0.99 ± 0.04	0.98 ± 0.05	0.99 ± 0.02	0.99 ± 0.02	0.99 ± 0.02	0.99 ± 0.02	0.99 ± 0.02	0.99 ± 0.02	0.99 ± 0.02
	GeoDE	0.98 ± 0.05	0.99 ± 0.02	0.99 ± 0.04	0.99 ± 0.02	0.99 ± 0.02	0.99 ± 0.02	0.99 ± 0.02	0.99 ± 0.02	0.99 ± 0.02
	ReliefF	0.98 ± 0.04	0.98 ± 0.04	0.99 ± 0.02	0.99 ± 0.02	0.99 ± 0.02	0.99 ± 0.02	0.99 ± 0.04	0.99 ± 0.04	0.99 ± 0.02
	Info_Gain	0.99 ± 0.02	0.99 ± 0.02	0.99 ± 0.02	0.99 ± 0.02	0.99 ± 0.02	0.99 ± 0.02	0.99 ± 0.02	0.99 ± 0.02	0.99 ± 0.02
	Func_Pert	0.98 ± 0.04	0.99 ± 0.02	0.99 ± 0.02	0.99 ± 0.02	0.99 ± 0.02	0.99 ± 0.02	0.99 ± 0.02	0.99 ± 0.02	0.99 ± 0.02
	EFSIS	0.99 ± 0.04	0.99 ± 0.02	0.99 ± 0.02	0.99 ± 0.02	0.99 ± 0.02	0.99 ± 0.02	0.99 ± 0.02	0.99 ± 0.02	0.99 ± 0.02
ProstateSingh	SAM	0.95 ± 0.08	0.95 ± 0.08	0.95 ± 0.08	0.96 ± 0.06	0.96 ± 0.07	0.95 ± 0.07	0.96 ± 0.07	0.96 ± 0.07	0.96 ± 0.07
	GeoDE	0.90 ± 0.15	0.93 ± 0.09	0.94 ± 0.09	0.95 ± 0.08	0.95 ± 0.09	0.94 ± 0.08	0.95 ± 0.06	0.95 ± 0.06	0.96 ± 0.06
	ReliefF	0.95 ± 0.09	0.94 ± 0.10	0.95 ± 0.08	0.96 ± 0.06	0.94 ± 0.09	0.96 ± 0.06	0.96 ± 0.06	0.95 ± 0.08	0.95 ± 0.08
	Info_Gain	0.94 ± 0.11	0.94 ± 0.10	0.94 ± 0.10	0.95 ± 0.09	0.95 ± 0.09	0.96 ± 0.07	0.97 ± 0.06	0.97 ± 0.06	0.96 ± 0.08
	Func_Pert	0.93 ± 0.10	0.95 ± 0.09	0.94 ± 0.10	0.96 ± 0.07	0.96 ± 0.07	0.96 ± 0.07	0.96 ± 0.08	0.96 ± 0.08	0.94 ± 0.09
	EFSIS	0.94 ± 0.10	0.94 ± 0.10	0.94 ± 0.09	0.95 ± 0.09	0.95 ± 0.08	0.97 ± 0.06	0.96 ± 0.08	0.95 ± 0.09	0.94 ± 0.09

Table 2. Prediction performance (Mean AUC ± standard deviation)

The best individual method is marked in green, and the ones that are significantly (p-value < 0.05, paired t-test) worse than the best individual method are marked in red.

- Some “best” individual methods in one dataset can be the “worst” in another one.
- Ensemble methods are worse than the best individual ones only in 3/54 tests.

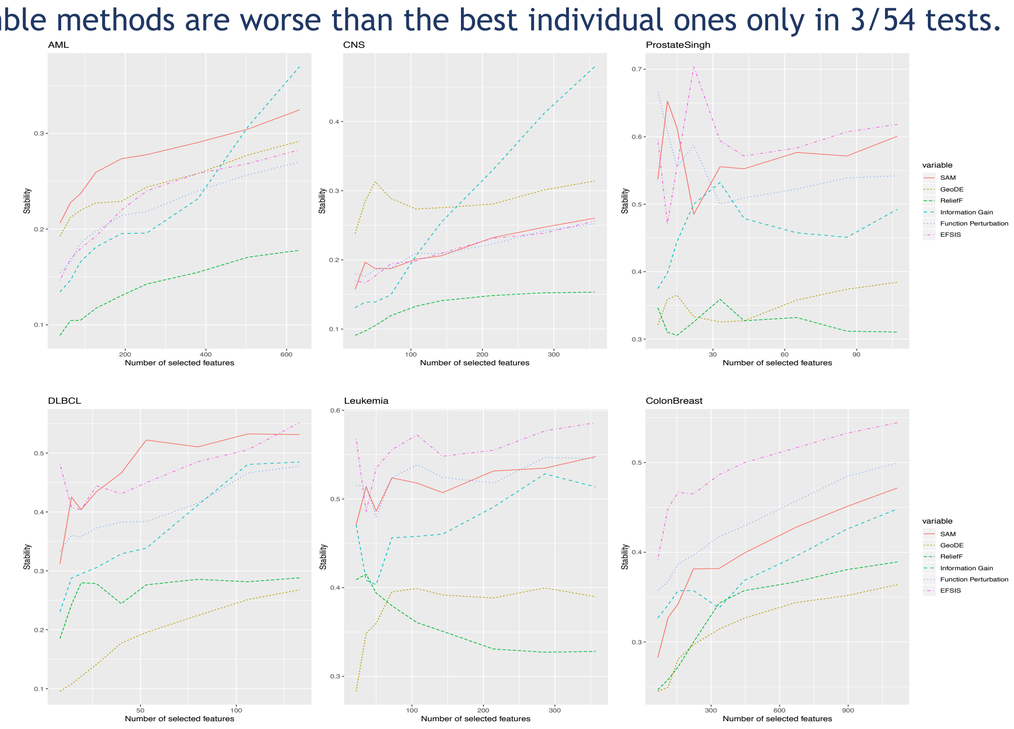

Figure 3. Stability performance

- EFSIS can improve the stability of function perturbation in most cases.
- EFSIS either has the best performance or performs moderately compared with all the individual methods.

References

[1] Breitling, Rainer, et al. "Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments." FEBS letters 573.1-3 (2004): 83-92.
[2] Tusher, V., et al. "Significance analysis of microarrays applied to the ionizing radiation response." Proceedings of the National Academy of Sciences 98.9 (2001): 5116-5121.
[3] Clark, Neil R., et al. "The characteristic direction: a geometrical approach to identify differentially expressed genes." BMC bioinformatics 15.1 (2014): 79.
[4] Kononenko, I. (1994). Estimating attributes: Analysis and extensions of RELIEF (pp. 171–182). Springer, Berlin, Heidelberg.
[5] Witten, I. H., et al. (2016). Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann.
[6] Yagi, T., et al. "Identification of a gene expression signature associated with pediatric AML prognosis." Blood102.5 (2003): 1849-1856.
[7] Pomeroy, Scott L., et al. "Prediction of central nervous system embryonal tumour outcome based on gene expression." Nature 415.6870 (2002): 436.
[8] Chowdary D., et al. Prognostic gene expression signatures can be measured in tissues collected in RNAlater preservative. J Mol Diag. 2006;8(1):31–9.
[9] Shipp, Margaret A., et al. "Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning." Nature medicine 8.1 (2002): 68.
[10] Golub, Todd R., et al. "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring." science 286.5439 (1999): 531-537.
[11] Singh, D., et al. "Gene expression correlates of clinical prostate cancer behavior." Cancer cell 1.2 (2002): 203-209.

decoding the systems toxicology of atlantic cod
The Research Council of Norway
CENTRE FOR DIGITAL LIFE NORWAY