Stability of Ensemble Feature Selection on High-Dimension and Low-Sample Size Data
Influence of the Aggregation Method

David Dernoncourt[1,2,3], Blaise Hanczar[4] and Jean-Daniel Zucker[1,2,3,5]

[1]*Institut National de la Santé et de la Recherche Médicale, U872, Nutriomique, Équipe 7,
Centre de Recherches des Cordeliers, Paris 75006, France*

[2]*Université Pierre et Marie-Curie - Paris 6, Nutriomique, 15 rue de l'École de Médecine, Paris 75006, France*

[3]*Institute of Cardiometabolism and Nutrition, Assistance Publique-Hôpitaux de Paris, CRNH-Île de France,
Pitié-Salpêtrière, Boulevard de l'Hôpital, Paris 75013, France*

[4]*LIPADE, Université Paris Descartes, 45 rue des Saint-Pères, Paris, F-75006, France*

[5]*Institut de Recherche pour le Développement, IRD, UMI 209, UMMISCO, France Nord, F-93143, Bondy, France*

Keywords: Feature Selection, Stability, Ensemble, Small Sample.

Abstract: Feature selection is an important step when building a classifier. However, the feature selection tends to be unstable on high-dimension and small-sample size data. This instability reduces the usefulness of selected features for knowledge discovery: if the selected feature subset is not robust, domain experts can have little trust that they are relevant. A growing number of studies deal with feature selection stability. Based on the idea that ensemble methods are commonly used to improve classifiers accuracy and stability, some works focused on the stability of ensemble feature selection methods. So far, they obtained mixed results, and as far as we know no study extensively studied how the choice of the aggregation method influences the stability of ensemble feature selection. This is what we study in this preliminary work. We first present some aggregation methods, then we study the stability of ensemble feature selection based on them, on both artificial and real data, as well as the resulting classification performance.

1 INTRODUCTION

Feature selection is a critical step of the supervised classification procedure and specially in the small-sample high dimension settings. Small-sample high dimension settings refers to problems where the number of features is higher than the number of examples. This kind of problem is increasingly frequent, especially in bioinformatics with the massive production of ”omics” data. In this context the learning algorithms met several problems called the curse of dimensionality (Simon, 2003). In high dimension, finding the actual informative features becomes more difficult and the risk of overfiting strongly increases. The consequence is the worse generalization performance of the classifiers (Jain and Chandrasekaran, 1982). Secondly, very high dimension alone is an issue itself, as classifiers frequently do not scale well to huge numbers of features, leading to increased computation times. Thirdly, a classifier based on a small subset of genes will be easier and less expensive to use in practice. Moreover, a classifier based on a high number of features will not be easily interpretable. The point problem is about the robustness of the selected features. To obtain a confident classifier, the selected subset has to be stable. To deal with all of these problems a feature selection is necessary in order to reduce dimensionality of the data.

Feature selection refers to the process of removing irrelevant or redundant features (in our context, genes) from the original set of features $\mathcal{F} = \{f_1, f_2, ..., f_{|\mathcal{F}|=D}\}$, so as to retain a subset $S \subset \mathcal{F}$ containing only informative features useful for classification (Liu et al., 2010). Beyond classification performance, the other main objective of the gene selection is to obtain a reliable and robust list of predictive genes: a gene which is regularly selected over several datasets dealing with the same problem – or at least over various random subsamples of the same dataset – is more likely to be really relevant, and is of greater interest to domain experts willing to use the classification results for knowledge discovery purposes. In

Dernoncourt D., Hanczar B. and Zucker J..
Stability of Ensemble Feature Selection on High-Dimension and Low-Sample Size Data - Influence of the Aggregation Method.
DOI: 10.5220/0004922203250330
In *Proceedings of the 3rd International Conference on Pattern Recognition Applications and Methods* (ICPRAM-2014), pages 325-330
ISBN: 978-989-758-018-5

practice, this is generally not the case, and many studies on feature selection stability emphasized the difficulty to obtain a reproducible gene signature on high-dimension small-sample data (Ein-Dor et al., 2006; Haury et al., 2011). Stability of feature selection strongly depends on the N/D ratio and problem difficulty. For example in one previous work we have shown that on a simple two Gaussian classes problem with 1000 features, a t-test based selection has a probability of more than 0.95 to select the very informative features when $N = 1000$, but this probability falls to less than 0.2 when $N = 50$ (Dernoncourt et al., 2014).

A growing number of studies have been dealing with feature selection stability, be it either to introduce stability measures (Kuncheva, 2007; Somol and Novovičová, 2010), to compare the stability of existing methods (Haury et al., 2011), and/or to propose innovative, stability-focused feature selection methods (Saeys et al., 2008; Han and Yu, 2012). Some works, such as (Somol et al., 2009; Abeel et al., 2010; Haury et al., 2011), also investigated the stability of ensemble feature selection methods. The main idea behind ensemble methods is to produce several individual feature selections and combine them in order to obtain a selection that outperforms every one of them. This idea is based on the concept of ”wisdom of the crowd”, which states that *under certain controlled conditions, the aggregation of information from several sources, results in decisions that are often superior to those that could have been made by any single individual - even experts* (Surowiecki, 2004; Rokach, 2010). In practice, ensemble methods have been widely applied to classifiers and since they can improve classifiers accuracy and stability, we can suppose that they should provide similar benefits to feature selection techniques (Yang et al., 2010). So far, works on ensemble feature selection have mainly focused on classification accuracy, showing accuracy gains (or losses) to be problem-dependent (Han et al., 2013) and filter-dependent (Wald et al., 2013). Works which also studied stability have obtained mixed results too, leaving the general impression that both stability and accuracy gains (or losses) from ensemble methods are problem-dependent (Saeys et al., 2008). However, those works have often measured stability over overlapping resamplings, which strongly increases the measured stability (Haury et al., 2011), and might also impact stability variations, and as far as we know no study extensively studied how the choice of the aggregation method influences the stability of ensemble feature selection methods.

In this preliminary work, we start investigating how ensemble methods improve feature selection (FS) stability, with a focus on the impact of the aggregation method. We first briefly present the stability measures and the three ensemble aggregation methods we used. Then we perform an empirical analysis of feature selection stability on both artificial and real microarray datasets.

2 ENSEMBLE FEATURE SELECTION

Creating an ensemble feature selection can be divided into two steps. The first step is to create a set of diverse feature selectors. The second step is to aggregate them.

2.1 Diversity Generation

The diversity of feature selectors is a crucial condition for obtaining a ”wise crowd” (Surowiecki, 2004), necessary for an efficient ensemble. It can be obtained via different methods such as:

- manipulating the training sample: typically, resampling the training set so as to perform each FS of the ensemble on a different training set,
- manipulating the FS method: for instance, use different parameters for each FS, if the FS method has parameters,
- partitioning the search space: each FS is performed on a different search space, for instance, random forest learns each tree on a random, different, small subset of features,
- hybridization: use several FS methods in the ensemble,

or a combination of those (Rokach, 2010). In this paper, we focused on manipulating the training sample, which is the most commonly used method, and obtained diversity by bootstrapping the training samples $B = 40$ times, based on previous works that showed that ensemble FS doesn’t improve much when increasing the amount of bootstrap samples further than 40 (Abeel et al., 2010) or even 20 (Saeys et al., 2008).

2.2 Aggregation

We tested the following methods of aggregation:

- Average score: on each bootstrap sample, the FS method outputs a score $s_{f_i,j}$ for each gene f_i. We simply average the score of a gene over the bootstrap samples in order to obtain the ensemble score W_{f_i}:

$$W_{f_i} = \frac{\sum_{j=1}^{B} s_{f_i,j}}{B} \quad (1)$$

- Average rank (Abeel et al., 2010): on each bootstrap sample, the score of each gene $s_{f_i,j}$ is converted into a rank $r_{f_i,j}$. Then we average the rank of a gene over the bootstrap samples in order to obtain the ensemble score:

$$W_{f_i} = \frac{\sum_{j=1}^{B} r_{f_i,j}}{B} \qquad (2)$$

- Stability selection (Meinshausen and Bhlmann, 2010): the ensemble score W_{f_i} of each gene is obtained by measuring how often the gene ranks in the top d of each bootstrap sample:

$$W_{f_i} = \frac{\sum_{j=1}^{B} I(f_i,j)}{B} \qquad (3)$$

  where $I(f_i,j) = 1$ if gene f_i is ranked in the top d on FS performed on the j^{th} bootstrap sample, and $I(f_i,j) = 0$ otherwise.

Finally, for each aggregation strategy, the d genes with the largest score W_{f_i} are retained as the final ensemble selection.

3 EXPERIMENTAL DESIGN

We performed experiments on both artificial and real data in order to assess the impact of ensemble methods on the stability of feature selection. We used four base FS methods: t-score, random forest, recursive feature elimination using support vector machines (SVM-RFE), and mutual information. For each of those methods, feature selection was performed with and without the ensemble aggregation methods described in the previous section. We then measured the stability of the feature selection on strictly non-overlapping sets (different generated sets on artificial data, resamplings on real data), and the average classification error rate obtained with a linear discriminant analysis (LDA) classifier on the test sets.

3.1 Stability Measure

Many different measures have been described to measure stability. As our main stability measure, we chose to use the relative weighted consistency (Somol and Novovičová, 2010), CW_{rel}, which evaluates how frequently each feature is selected, among the features which have been selected at least once. We chose it because it ignores the stability of non-selected features (which would artificially increase the stability measure on large datasets where many features are irrelevant and easy to exclude), and because it's adjusted to take into account the proportion of overlapping features due to chance.

Let $\mathcal{S} = \{S_1, S_2, ..., S_\omega\}$ be a system of ω gene subsets obtained from ω runs of the feature selection routine on different samplings, $\Omega = \sum_{i=1}^{\omega} |S_i|$ be the total number of occurrences of any gene in $\mathcal{S}$ and F_f be the number of occurrences of gene $f \in \mathcal{F}$ in system $\mathcal{S}$. The weighted consistency CW was defined as follow:

$$CW(\mathcal{S}) = \sum_{f \in X} \frac{F_f}{\Omega} \cdot \frac{F_f - 1}{\omega - 1} \qquad (4)$$

and CW_{rel} is obtained by adjusting CW on its minimal and maximal possible values CW_{min} and CW_{max}:

$$CW_{rel}(\mathcal{S},\mathcal{F}) = \frac{CW(\mathcal{S}) - CW_{min}(\Omega,\omega,\mathcal{F})}{CW_{max}(\Omega,\omega) - CW_{min}(\Omega,\omega,\mathcal{F})} \qquad (5)$$

3.2 Artificial Data

We used three different artificial data structures, all based on a two-classes Gaussian model with $D = 1000$ genes. Each of the two classes follows a normal distribution defined respectively by $\mathcal{N}(\mu, I)$ and $\mathcal{N}(-\mu, \Sigma)$, where μ is a vector of means such that $|\mu| = D$ and Σ is the covariance matrix.

In the first data structure, *NC100*, all genes are independent (Σ is the identity matrix) and the elements μ_i of μ consisted of $d = 100$ elements $\mu_i = 1$ (genes useful for classification) and $D - d = 900$ elements $\mu_i = 0$ (noise genes). Then μ was scaled down so that $\mathcal{F}$ would yield a specified Bayes error ($\varepsilon_{Bayes} = 0.10$).

In the second data structure, *NC*, all genes are independent and the elements μ_i of μ were drawn from a triangular distribution with a lower limit and mode equal to 0 (probability density function: $f(x) = 2 - 2x$ for $x \in [0;1]$). In order to obtain a more realistic probability density, we then raised μ to a power of $\gamma = 2$, similarly to the method used in (Dernoncourt et al., 2014). Again, μ was scaled down so that $\mathcal{F}$ would yield $\varepsilon_{Bayes} = 0.10$.

In the third data structure, *CB*, we added correlations within blocks of ten genes, by using the covariance matrix

$$\Sigma = \left| \begin{array}{cccc} \Sigma_1 & 0 & \cdots & 0 \\ 0 & \Sigma_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \Sigma_{100} \end{array} \right|, \qquad (6)$$

where Σ is a block diagonal matrix and Σ_i is a 10×10 square matrix with elements 1 along its diagonal and 0.5 off its diagonal, similarly to the method used in (Han and Yu, 2012). We used the same μ as in the *NC* dataset.

From these models, 50 training sets were generated, on which the FS methods were performed and an LDA classifier was trained. Each classifier was then applied to a test set consisting of 10000 instances to estimate its error rate.

Table 1: Microarray datasets.

Name	N	D	Source
Colon cancer	62	2000	Alon et al., 1999
Leukemia	72	7129	Golub et al., 1999
Breast cancer Pawitan	159	8112	Pawitan et al., 2005
Lung cancer	203	2000	Bhattacharjee et al., 2001
Breast cancer Vijver	294	2000	van de Vijver et al., 2002

3.3 Real Data

We experimented with five publicly available microarray datasets, listed in Table 1. For each dataset, 50 training sets were generated by randomly drawing half instances from the dataset (without replacement). For each of them, feature selection was performed and a classifier was trained (using the same methods as with the artificial data). Each classifier was then applied on a test set consisting of the samples not included in the corresponding training set. Stability of the feature selection was measured within each pair of training and test sets, so as to have no overlap. The final measure of stability corresponds to the average of those 50 measures.

4 RESULTS

4.1 Artificial Data

Table 2 presents the stability of feature selections and the error rate of resulting classifiers on the artificial datasets. In general, we observed that ensemble feature selection provides similarly or more stable results than non-ensemble (single) feature selection, and a similar or lower error rate.

T-score obtained the highest stability and lowest classification error rate overall. It further benefited from its ensemble version, but only with the average score aggregation: the average rank and stability selection aggregation slightly degraded its stability and did not improve the classification error rate.

Compared to t-score, SVM-RFE had a slightly lower stability on non-correlated data and half the stability on correlated data, but with a similar error rate. It did not benefit from the ensemble version on non-correlated data, but stability was improved on correlated data, similarly with all aggregation types.

Random forest was the worst performer, with half the stability of t-score on all datasets, and a higher error rate on the uncorrelated datasets. It was however much more stable in its ensemble version, reaching similar stability levels and error rate as single SVM-RFE on non-correlated data, and a higher stability (and somewhat lower error rate) than ensemble SVM-RFE on correlated data. As with SVM-RFE, the aggregation method did not matter here.

4.2 Real Data

Table 3 presents the stability of feature selections and the error rate of resulting classifiers on the real microarray datasets. In general, similarly to the artificial data, we observed that ensemble feature selection provides similarly or more stable results than single feature selection, with the exception of t-score on the leukemia dataset. Unlike what we observed on artificial data though, the error rate was in some cases increased by the ensemble selection.

T-score obtained the highest stability on 2 out of 5 datasets. In 4 datasets, its ensemble version had an improved stability, but only with the average score aggregation, and in half cases (colon cancer and lung cancer datasets) at the cost of a largely increased error rate. On the leukemia dataset, ensemble methods reduced stability and increased error rate.

SVM-RFE obtained the highest stability on 3 out of 5 datasets. In all datasets, its ensemble version had an improved stability, the best improvement occurring with the average score aggregation, closely followed by stability selection. Aggregation by average rank did not perform as well: it provided the worse stability increase on the colon cancer and lung cancer dataset, and no stability increase or even a stability degradation on the other datasets. The error rate was generally unchanged by ensemble methods, except for a 10% increase in the colon cancer dataset (with any aggregation method) and a 20% decrease in the lung cancer dataset (aggregation by average rank only).

Random forest had the lowest stability on 3 out of 5 datasets, and tied with mutual information on the Vijver dataset. However it had a competitive error rate (best without ensemble on Vijver and leukemia datasets, best with ensemble on the colon cancer dataset), even though the differences in error rates between the different feature selection methods were generally small, except on the Pawitan dataset. Similarly to SVM-RFE, its stability was generally increased by ensemble methods. Aggregation by average score and stability selection improved stability

Table 2: Classification error rate and selection stability on the artificial datasets, with feature selection without ensemble and with three ensemble aggregation methods.

Artificial data type	Ensemble aggregation	t-score		SVM-RFE		Random forest	
		Error rate	CW_{rel}	Error rate	CW_{rel}	Error rate	CW_{rel}
NC100	Single	0.338	0.075	0.345	0.065	0.383	0.031
	Average score	0.290	0.135	0.339	0.071	0.347	0.059
	Average rank	0.340	0.072	0.340	0.072	0.347	0.061
	Stability selection	0.339	0.071	0.344	0.068	0.345	0.062
NC	Single	0.360	0.050	0.360	0.049	0.391	0.027
	Average score	0.320	0.071	0.364	0.045	0.365	0.042
	Average rank	0.361	0.045	0.364	0.044	0.363	0.045
	Stability selection	0.366	0.043	0.363	0.045	0.368	0.040
CB	Single	0.239	0.235	0.242	0.113	0.244	0.099
	Average score	0.223	0.265	0.238	0.152	0.232	0.193
	Average rank	0.235	0.226	0.238	0.154	0.230	0.195
	Stability selection	0.237	0.217	0.239	0.150	0.233	0.190

Table 3: Classification error rate and selection stability on the microarray datasets, with feature selection without ensemble and with three ensemble aggregation methods.

Data	Ensemble aggregation	t-score		SVM-RFE		Random forest		Mutual inf	
		Error	CW_{rel}	Error	CW_{rel}	Error	CW_{rel}	Error	CW_{rel}
Colon	Single	0.188	0.310	0.182	0.448	0.196	0.163	0.181	0.140
	Score	0.305	0.327	0.203	0.588	0.179	0.206	0.217	0.149
	Rank	0.215	0.277	0.203	0.494	0.199	0.163	0.209	0.149
	Stability	0.210	0.262	0.206	0.568	0.177	0.210	0.213	0.145
Leukemia	Single	0.049	0.322	0.044	0.525	0.042	0.220	0.050	0.263
	Score	0.117	0.315	0.051	0.581	0.043	0.265	0.052	0.300
	Rank	0.054	0.269	0.047	0.517	0.067	0.094	0.053	0.297
	Stability	0.058	0.246	0.047	0.565	0.046	0.269	0.049	0.294
Pawitan	Single	0.342	0.071	0.283	0.129	0.309	0.011	0.313	0.023
	Score	0.328	0.085	0.283	0.180	0.320	0.012	0.343	0.024
	Rank	0.344	0.065	0.298	0.095	0.324	0.015	0.335	0.024
	Stability	0.317	0.047	0.289	0.180	0.314	0.011	0.329	0.022
Lung	Single	0.054	0.515	0.084	0.377	0.060	0.342	0.061	0.372
	Score	0.076	0.536	0.083	0.498	0.064	0.398	0.063	0.379
	Rank	0.058	0.417	0.067	0.444	0.063	0.389	0.064	0.377
	Stability	0.058	0.445	0.082	0.487	0.063	0.394	0.064	0.376
Vijver	Single	0.382	0.254	0.377	0.159	0.359	0.078	0.360	0.077
	Score	0.364	0.345	0.373	0.221	0.368	0.107	0.357	0.091
	Rank	0.371	0.237	0.368	0.158	0.371	0.105	0.360	0.092
	Stability	0.374	0.237	0.376	0.215	0.371	0.106	0.359	0.088

equally and on all datasets, except for the Pawitan dataset. Aggregation by average rank was more inconsistent: similar improvements as the other aggregation methods on Vijver and lung cancer datasets, a better improvement on the Pawitan dataset, where random forest had a very low stability compared to the other methods, no improvement compared to single random forest on the colon cancer dataset, and an important degradation on the leukemia dataset.

Mutual information had the lowest stability on the colon cancer dataset, yet competitive error rates. The ensemble versions had a similar (colon cancer, Pawitan, and lung cancer datasets) or moderately improved (leukemia and Vijver datasets) stability, with no marked difference in favor of a specific aggregation method. Ensemble increased the error rate in the colon cancer and Pawitan dataset, again with no marked difference between aggregation methods.

5 CONCLUSIONS AND FUTURE WORK

In this work, we studied how ensemble feature selection methods influence the stability of the gene selection and, to a lesser extent, the error rate of the resulting classifier on microarray data and artificial data of similar dimension. We focused on the aggregation method used in the ensemble procedure, because that is an important aspect of the ensemble construction procedure which, to our knowledge, had scarcely been investigated in such a setting before. Similarly to (Haury et al., 2011), we found that average rank aggregation usually performed worse than the other aggregation methods. We also found that average score aggregation usually resulted, with a few exception, in the best stability, while stability selection aggregation was in-between.

We observed, in some cases, a trade-off between stability and error rate. Previous studies such as (Saeys et al., 2008) already suggested such a dataset-dependent trade-off between robustness and classification performance. Here, however, we find that the aggregation method can also play a role in this trade-off, since in some cases error rate and stability were differently affected by the different ensemble aggregation methods. This trade-off did not seem to apply to our artificial data, though: on them, a better error rate was systematically paired with a better stability. This difference could be due to structural differences with the real data (the latter probably presenting much more complex and numerous interactions), or to a lack of diversity in the artificial data, since the trade-off is not observed in all datasets. Nonetheless, we observed that in all cases, the most stable method without ensemble could be rendered more stable via ensemble, with or without a trade-off on the classification error rate.

As future work, we think it would be interesting to study or develop more aggregation methods, such as average weighted rank or score (giving a higher weight to higher scores or lower ranks). Weighted (exponential) rank reportedly performed better than average rank (Haury et al., 2011), so maybe such an improvement could be obtained by using the same method on scores. Hybridization of different base feature selection methods also seems to be an interesting area to explore, which will also require some specific work on the aggregation strategies.

REFERENCES

Abeel, T., Helleputte, T., Van de Peer, Y., Dupont, P., and Saeys, Y. (2010). Robust biomarker identification for cancer diagnosis with ensemble feature selection methods. *Bioinformatics*, 26(3):392–398.

Dernoncourt, D., Hanczar, B., and Zucker, J.-D. (2014). Analysis of feature selection stability on high dimension and small sample data. *Computational Statistics and Data Analysis*, 71(0):681 – 693.

Ein-Dor, L., Zuk, O., and Domany, E. (2006). Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer. *Proceedings of the National Academy of Sciences*, 103(15):5923–5928.

Han, Y., Yang, Y., and Zhou, X. (2013). Co-regularized ensemble for feature selection. In *Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence*, IJCAI'13, pages 1380–1386. AAAI Press.

Han, Y. and Yu, L. (2012). A variance reduction framework for stable feature selection. *Statistical Analysis and Data Mining*, 5(5):428–445.

Haury, A.-C., Gestraud, P., and Vert, J.-P. (2011). The influence of feature selection methods on accuracy, stability and interpretability of molecular signatures. *PLoS ONE*, 6(12):e28210.

Jain, A. K. and Chandrasekaran, B. (1982). 39 dimensionality and sample size considerations in pattern recognition practice. *Handbook of Statistics*, 2:835–855.

Kuncheva, L. I. (2007). A stability index for feature selection. In Devedzic, V., editor, *Artificial Intelligence and Applications*, pages 421–427. IASTED/ACTA Press.

Liu, H., Liu, L., and Zhang, H. (2010). Ensemble gene selection by grouping for microarray data classification. *Journal of biomedical informatics*, 43:81–87.

Meinshausen, N. and Bhlmann, P. (2010). Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(4):417–473.

Rokach, L. (2010). Ensemble-based classifiers. *Artificial Intelligence Review*, 33(1-2):1–39.

Saeys, Y., Abeel, T., and Peer, Y. (2008). Robust feature selection using ensemble feature selection techniques. In Daelemans, W., Goethals, B., and Morik, K., editors, *Machine Learning and Knowledge Discovery in Databases*, volume 5212 of *Lecture Notes in Computer Science*, pages 313–325. Springer Berlin Heidelberg.

Simon, R. (2003). Supervised analysis when the number of candidate features (p) greatly exceeds the number of cases (n). *SIGKDD Explor. Newsl.*, 5(2):31–36.

Somol, P., Grim, J., and Pudil, P. (2009). Criteria ensembles in feature selection. In Benediktsson, J. A., Kittler, J., and Roli, F., editors, *MCS*, volume 5519 of *Lecture Notes in Computer Science*, pages 304–313. Springer.

Somol, P. and Novovičová, J. (2010). Evaluating stability and comparing output of feature selectors that optimize feature subset cardinality. *IEEE Trans. Pattern Anal. Mach. Intell.*, 32(11):1921–1939.

Surowiecki, J. (2004). *The Wisdom of Crowds: Why the Many are Smarter Than the Few and how Collective Wisdom Shapes Business, Economies, Societies, and Nations*. Doubleday.

Wald, R., Khoshgoftaar, T. M., and Dittman, D. J. (2013). Ensemble gene selection versus single gene selection: Which is better? In Boonthum-Denecke, C. and Youngblood, G. M., editors, *FLAIRS Conference*. AAAI Press.

Yang, P., Hwa Yang, Y., B. Zhou, B., and Y. Zomaya, A. (2010). A review of ensemble methods in bioinformatics. *Current Bioinformatics*, 5(4):296–308.