Proceedings of the 2012 IEEE 12th International Conference on Bioinformatics
& Bioengineering (BIBE), Larnaca, Cyprus, 11-13 November 2012

Experimental Analysis of Feature Selection Stability for High-Dimension and Low-Sample Size Gene Expression Classification Task

David Dernoncourt*†, Blaise Hanczar‡ and Jean-Daniel Zucker*†§
*Institut National de la Santé et de la Recherche Médicale, U872, Nutriomique, Équipe 7,
Centre de Recherches des Cordeliers, 75006, Paris, France - Email: david.dernoncourt@crc.jussieu.fr
†Université Pierre et Marie-Curie - Paris 6, 75006, Paris, France
‡LIPADE, Université Paris Descartes, 45 rue des Saint-Pères, Paris, F-75006 France
§Institut de Recherche pour le Développement, IRD, UMI 209, UMMISCO, France Nord, F-93143, Bondy, France

***Abstract*—Gene selection is a crucial step when building a classifier from microarray or metagenomic data. As the number of observations is small, the gene selection tends to be unstable. It is common that two gene subsets, obtained from different datasets but dealing with the same classification problem, do not overlap significantly. Although it is a crucial problem, few works have been done on the selection stability. In this paper, we first present some stability quantification methods, then we study the variations of those measures with various parameters (dimensionality, sample size, feature distribution, selection threshold) on both artificial and real data, as well as the resulting classification performance. Feature selection was performed with t-test and classification with linear discriminant analysis. We point out a strong empiric correlation between the dimensionality/sample size ratio and selection instability.**

***Index Terms*—Feature selection, small sample, stability, dimensionality/sample size ratio.**

I. Introduction

High-throughput technologies have allowed the production of large genomic datasets: for instance, microarray data contain the simultaneous expression of tens of thousands of genes whereas NGS (Next Generation Sequencing) may reach several millions of genes. The use of supervised learning methods on these data makes classifiers predicting different medical parameters. This classification task may be very useful in the medical decision strategy, for instance, the classification of tumor types, the prediction of the clinical outcome [1], or for early disease detection [2]. We expect that in the next years, these classifiers based on genomic data will help physicians to take the right decision.

Although the microarray data contain the expression of several thousands of genes, the final classifiers should be based only on a small subset of genes. The first reason comes from the disproportion between the number of genes and samples of microarray datasets. Due to cost problems, microarray datasets usually contain few patients (at most a few hundreds), leading to what is commonly known as $N << D$ problems: classification tasks in which the number of features D is much larger than the number of samples N. High dimensionality and small sample size both increase the risk of overfitting and decrease the accuracy of classifiers [3]. The second reason is practical: it is easier and less expensive to use a classifier based on a small subset of genes than on several thousands. To deal with these problems, a gene selection is applied on the data before classifier construction in order to reduce dimensionality.

Feature selection refers to the process of removing irrelevant or redundant features (in our context, genes) from the original set of genes, so as to retain a subset containing only informative genes useful for classification. Feature selection methods can be broken down into three categories: filter, wrapper and embedded methods. It is generally agreed that wrapper or embedded methods should be preferred if it is technically feasible [4]. However, on very high dimensional data, filters remain the method of choice for tractability reasons.

Beyond classification performance, the other main objective of the gene selection is to obtain a reliable and robust list of predictive genes (signature). To validate clinically a diagnosis system based on a classifier, the results and the gene selection must be reproducible. It is therefore crucial that the gene selection is stable, i.e. for a given classification task an accurate gene selection identified on a dataset must be accurate for the other datasets. Several groups have published signatures and reported good classification performance, but unfortunately the different signatures obtained, for the same classification task, differed widely. The number of genes shared by different signatures is not significantly higher than the overlap between random selections. For instance, in [5], five classification tasks dealing with a similar problem (breast cancer prognosis prediction from gene expression data) were performed on five different datasets, leading to very little overlap of the selected genes. Several other studies, such as [6] and [7], emphasized the difficulty to obtain a reproducible gene signature on small-sample microarray datasets. The lack of stability of gene selection is a blocking point in the development of classifiers. As long as this stability problem of gene selection is not solved, the classifiers based on genomic data cannot get from lab's experiments to medical applications in hospital. Some works proposed feature selection methods improving selection stability [8], [9]. However, none has been tested on several

978-1-4673-4358-9/12/$31.00 ©2012 IEEE

$N << D$ datasets from the same classification task.

In this paper, we show that an acceptable stability of the gene selection cannot be reached on most microarray datasets. We investigate the behaviour of the feature selection stability and its impact on the classifiers. We present two complementary and unbiased stability measures, then we perform an empirical analysis of feature selection stability on both artificial and real microarray datasets. This allowed us to shed more light on the influence of many dataset characteristics (number of examples, selection threshold, number of variables, variable distribution) on feature selection stability. Our simulations point out the lack of stability on small sample and high dimensional data, with a notably strong relationship between the N/D ratio and feature selection stability, and the resulting classification performances. We provide an empirical lower bound for the number of samples needed to reach a given level of stability.

II. Stability measures

The stability of a feature selection method can be defined as the modestness of changes in the set of selected genes when there are slight changes in training data. To evaluate it, many different measures have already been described. We chose to use four stability measures, which we will present according to the taxonomy presented by Somol and Novovičová [10].

A. Relative weighted consistency, an unbiased feature-focused measure

Among the stability measures sorted in the above-mentioned taxonomy, only one was both *selection-registering* and *subset-size-unbiased*: the relative weighted consistency CW_{rel} [10]. It is based on a *subset-size-biased* measure, the weighted consistency CW, corrected to be actually bounded by $[0;1]$ no matter what the proportion of selected genes is. A value of 0 indicates the lowest possible stability, while a value of 1 indicates the highest possible stability, i.e., if all feature subsets have the same cardinality, all subsets are identical.

Let $\mathcal{F} = \{f_1, f_2, ..., f_{|\mathcal{F}|=D}\}$ be the set of features and $\mathcal{S} = \{S_1, S_2, ..., S_\omega\}$ 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 $\mathcal{S}$. CW was defined as follow:

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

and CW_{rel} was then derived by adjusting CW on its minimal and maximal possible values CW_{min} and CW_{max}:

$$CW_{rel}(\mathcal{S}, \mathcal{F}) = \frac{CW(\mathcal{S}) - CW_{min}(\Omega, \omega, \mathcal{F})}{CW_{max}(\Omega, \omega) - CW_{min}(\Omega, \omega, \mathcal{F})} \tag{2}$$

B. Partially adjusted average Tanimoto index, an unbiased subset-focused measure

CW_{rel} is a *feature-focused* measure, so we looked for a *subset-focused* measure to complement it. Kuncheva's stability index [11] and the stability measure defined in [12] are both *subset-focused*, but they can only be used on subsets of equal cardinality. We retained the Average Tanimoto Index ATI, also introduced in [10]. ATI is a generalization based on Kalousis's similarity measure S_S between two sets S_i and S_j [13]:

$$S_S(S_i, S_j) = \frac{|S_i \cap S_j|}{|S_i \cup S_j|} \tag{3}$$

This similarity index is computed over all subset pairs, then averaged:

$$ATI(\mathcal{S}) = \frac{2}{\omega(\omega-1)} \sum_{i=1}^{\omega-1} \sum_{j=i+1}^{\omega} S_S(S_i, S_j) \tag{4}$$

ATI is *subset-focused* and *selection-registering*, but it is also *subset-size-biased.* We propose a correction of this index, the partially adjusted average Tanimoto index, defined as follow:

$$ATI_{PA}(\mathcal{S}) = Max(\frac{ATI(\mathcal{S}) - ATI_{exp}(\mathcal{S})}{ATI_{max}(\mathcal{S}) - ATI_{exp}(\mathcal{S})}, 0) \tag{5}$$

where ATI_{max} is the maximal possible value of ATI and ATI_{exp} is the expected value of ATI when genes subsets are randomly defined. Because we will use a feature selection method which outputs a subset of predefined size, $ATI_{max} = CW_{max} = 1$ (when all genes subsets are identical). To obtain ATI_{exp}, we used an experimentally-determined estimation, computed as a function of the proportion of selected features. It should be noted that the correction we performed in ATI_{PA} slightly differs from the one performed in CW_{rel}: CW_{rel} is adjusted on the smallest possible value, while ATI_{PA} is adjusted on the expected value. The max operator ensures that ATI_{PA} is within the $[0;1]$ interval and not negative as it could happen for the first argument of the max if the stability happens to be worse than random.

C. Correlation-based measures

Both ATI and CW focus on the stability of selected genes. This aspect is important for knowledge discovery but, for the purpose of evaluating feature selection methods, the stability of the ranking score over all genes may be an interesting information, too. *Selection-exclusion-registering* measures will be too biased when the proportion of excluded genes is too high. Correlations of features scores and ranks, on the other hand, provide a more balanced overview. However the latter will be penalized when lots of genes have a similar relevance, which occurs for example when a lot of genes are equally irrelevant in a very high-dimensional dataset. So, we used the average score (or weight) correlation $\overline{S_W}$ and the average rank correlation $\overline{S_R}$, as described in [13].

III. Experimental design

We performed a set of experiments on both artificial and real data in order to assess the impact of different dataset parameters on the gene selection. The investigated parameters are: the number of samples (N), the number of features (D), the number of selected features for the construction of the classifier (d) and the distribution of feature discrimination power (controlled by a parameter γ introduced in section III-A).

A. Artificial data

We generated artificial data that have the main characteristics of microarray data, i.e. few samples, a large number of features a large proportion of which is useless w.r.t. the classification task. The artificial data are based on a two-classes Gaussian model. Each of the two classes follows a normal distribution defined respectively by $\mathcal{N}(\mu, I)$ and $\mathcal{N}(-\mu, I)$. The values μ_i represent the discrimination power for each feature i, i.e. the higher μ_i, the more information the feature contains for the classification. 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 $x \in [0; 1]$). To obtain various shapes of strictly decreasing probability densities, simulating varying feature dispersion and relevance, we then raised μ to a power of γ ($\mu_i = \mu_i^\gamma, \gamma \in [1; 10]$). Finally, μ was scaled down so that either $\mathcal{F}$ would yield a specified Bayes error (ϵ_{Bayes}) or that the largest μ_i had a specific value μ_{imax}. In our experiments we chose $\epsilon_{Bayes} = 0.10$ or $\mu_{imax} = 0.15$. From this model, training and test sets are generated with 50 to 10000 features. Training sets contains 25 to 10000 examples so that the N/D ratio goes from 0.0025 to 200, exceeding the range of N/D seen in real datasets.

The score used to rank features on the training data was the absolute value of the t-score. Then the top d features with the highest score were selected. We chose the t-test because it should perform optimally on independent and normally distributed features such as our artificial data. For various combinations of parameters N, D, d and γ, and for various values of N and D while keeping the N/D ratio constant, 100 training sets were generated. For each of them, feature selection was performed and a linear discriminant analysis (LDA) classifier was trained. Each classifier was then applied on a test set consisting of 10000 examples. Besides the stability measures described in section II, we measured the frequency with which each feature was selected.

B. Real data

We experimented with three publicly available microarray datasets, related to lung cancer [14] ($D = 2000$, $N = 203$, $N/D \approx 0.10$) leukemia [15] ($D = 7129$, $N = 72$, $N/D \approx 0.01$) and breast cancer [16] ($D = 2000$, $N = 295$, $N/D \approx 0.15$). For each datasets, for different values of N, 100 training sets were generated by randomly drawing examples 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. We measured the stability of the feature selection accross the training sets (two different measures: stability accross all training sets at once, and the average of stabilities within each training set - test set pair) and the average classification error rate.

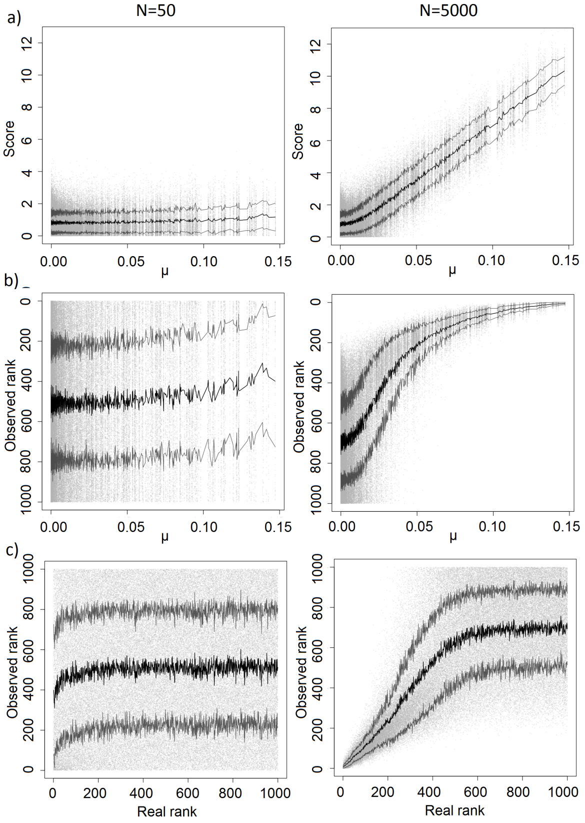

Fig. 1. On artificial data, with $D = 1000$, $d = 100$, $\gamma = 2$ and $N = 50$ (left) or $N = 5000$ (right). a) observed score (in absolute value) given real μ_i, b) observed rank given real μ_i, c) observed rank given real rank. One point per feature and per training set, the black curve is the average per feature, the grey curves the average ± standard deviation.

IV. Results

A. Artificial data

In this set of simulations, we present the performance and stability of the feature selection depending on dataset parameters.

Figure 1 provides an intuitive overview of feature scoring stability in two extreme settings: one with a very small sample ($N = 50$, left column), the other one with a large sample ($N = 5000$, right column). In the small sample case, feature scores (Figure 1a)) do not vary much with feature μ_i, and even though the most relevant features have a slightly higher score than the least relevant ones on average, their scores vary approximately on the same range. This contrasts with the large sample case, where the most relevant features have scores in the [8;12] range, far away from the least relevant ones, which stay in the [0;3] range and are thus easy to tell apart. The correlation between feature scores and μ_i decreases with N.

The resulting ranks reflect the inconsistency of the scores. Figure 1b) represents observed feature ranks given feature μ_i, Figure 1c) provides a slightly different visualization: observed

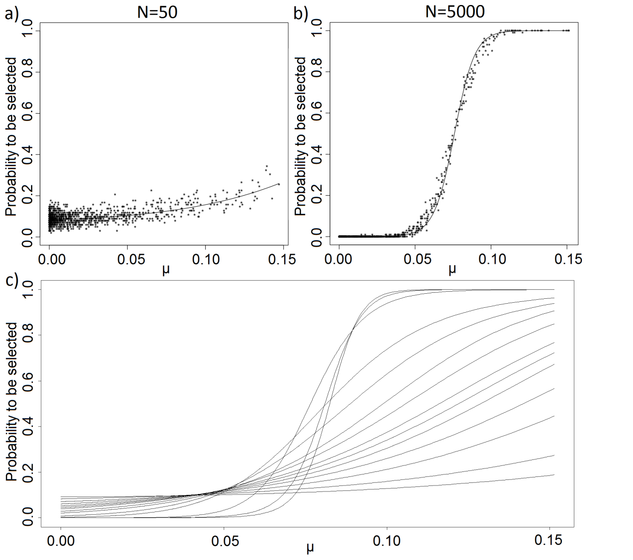

Fig. 2. Observed probability for a feature to be selected given its real μ_i (i.e. its increase of relevance). On artificial data, with $D = 1000$, $d = 100$ and $\gamma = 2$. a) $N = 50$ b) $N = 5000$, one point per feature and the curve was obtained via logistic regression. c) N varying from 25 (curve with the lowest value at $\mu_i = 0.15$) to 10000 (first curve to reach 1). As the sample size grows, the logistic shape increasingly stands out.

feature ranks (computed from scores) given true feature ranks (computed from μ_i). Due to the way our model was conceived, the least relevant features can be considered as noise, even though technically they do have some very tiny relevance. So having the worst features poorly ranked among each other is not a bad result. However, in the small sample case, even the true best features only have a slightly better average rank than the other features. For over 90% of the remaining features, the assigned rank is pure noise, as illustrated by scatter plot and standard deviation lines (gray curves) on figure 1. In the large sample case, the true best features are ranked much more accurately, even though some noise remains among them, and only the worst half of features are assigned to a mostly noisy rank. The results show that in small-sample data there is no correlation between the feature score and ranking obtained from the selection methods and the actual quality of the features.

From our simulations, we computed empirically the probability of each feature to be selected. Figure 2 presents the evolution of this probability given μ_i. We can see that in the small sample case (Figure 2a)), the probability for the most relevant features to actually be selected does not reach 35%, while even the least relevant features have a non negligible probability of being selected. In the large sample case (Figure 2b)), the selection is much more accurate: all features with $\mu_i > 0.10$ have a probability to be selected close to 1 and all features with $\mu_i < 0.05$ are almost never selected. Figure 2c) shows the evolution of the regression curve from $N = 25$ to $N = 10000$: as the sample size increases, the logistic shape increasingly stands out, illustrating how the selection progressively becomes more accurate. But only when the sample size reaches around 1000 observations is the feature selection algorithm able to select the most relevant features with a good sensitivity. In small sample data, the probability to reliably select good features is therefore very low.

Figure 3 presents the evolution of stability measures under varying dataset parameters. The stability is much influenced by the sample size N, with stability measures close to zero when the sample size is around 100 and increasing a lot when additional samples are added to the training set, up to 0.6+ for AIT_{PA} and almost 0.9 for $\overline{S_W}$. It is also much influenced by the total number of variables D, with fairly high values (0.4 to 0.6) when the dataset only contains 100 samples and 50 variables, quickly reaching close to zero with so few as 1000 variables.

To a lesser extent, stability is also influenced by the selection threshold d. In this case, CW_{rel} is minimal when we select very few variables, then it increases to reach a maximum when we select around 150-180 variables, finally it slowly but regularly decreases as we add more unreliable variables. The shape of this curve illustrates the difficulty to reliably identify even the most relevant variables: trying to keep just the 2 best features will yield highly unstable results. Trying to keep the 50 best features will include maybe the 5 or 10 best features with a very high reliability, leading to a higher stability even though the rest of the selection is not as stable. Note that in this setting, obviously $\overline{S_W}$ and $\overline{S_R}$ do not vary, as they do not take into account the fact that a feature was selected or not.

Variable distribution γ also has some influence on stability: stability measures are minimal when variables are distributed on the triangular distribution, and increase with γ, but following different patterns. $\overline{S_W}$ always increases: this measure is not penalized by ranking difficulties or by instability in the final selection, and only benefits from variables taking extreme values: variables with initial μ_i close to zero do not really lose much score correlation when they get squished even closer to zero, while variables with higher μ_i do benefit from getting more isolated farther away from zero. $\overline{S_R}$ increases at first, but then starts decreasing after reaching a maximum at $\gamma = 5$: this measure first benefits from the increased dispersion of variables with high or intermediate μ_i, but at some point this effect is overcome by the increased difficulty to rank variables with intermediate μ_i (because we kept a constant, realistic Bayes error, the more we streched the distribution the harder intermediately relevant variables became to identify), which eventually get too close to zero. CW_{rel} and ATI_{PA}, which perform their selection based on a cutoff in the ranks, evolve as a consequence of $\overline{S_R}$. However their decrease is somewhat delayed because they are only affected by the top d rankings. It is likely that a subset-size optimizing feature selection method would see a higher influence of data distribution over selection stability, because it would probably drop the decreasingly relevant variables while keeping the ones increasingly easier to identify.

Figure 4 shows the stability CW_{rel} as a function of the number of training examples for a constant N/D ratio. The different curves correspond to different N/D ratios (from 0.01

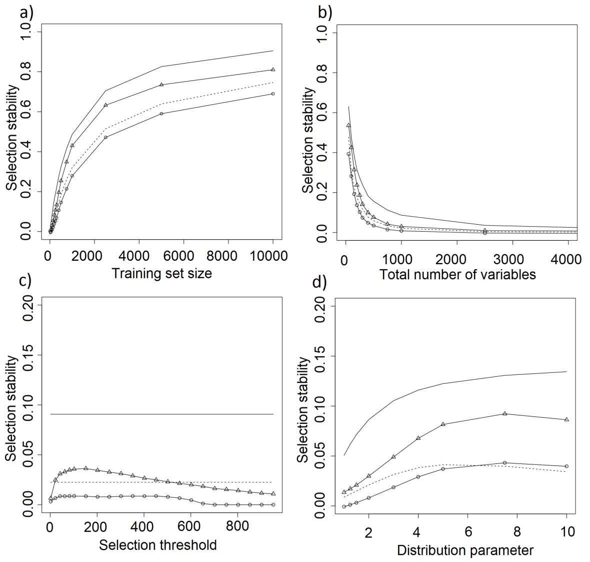

Fig. 3. Evolution of stability measures CW_{rel} (triangles), ATI_{PA} (circles), $\overline{S_W}$ (continuous line) and $\overline{S_R}$ (dashes) given: a) $N \in [25; 10000]$ b) $D \in [50; 10000]$ c) $d \in [2; 1000]$ and d) $\gamma \in [1; 10]$. When they were not the one being iterated on, parameter values were: $N = 1000$, $D = 100$, $d = D{\cdot}10\%$, $\gamma = 2$.

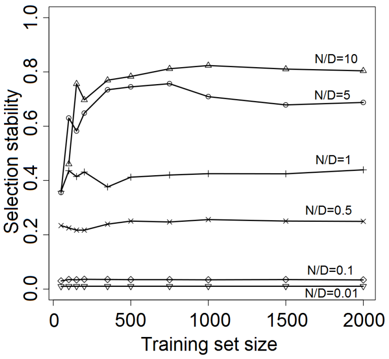

Fig. 4. Evolution of CW_{rel} with the number of training examples for a constant N/D ratio. The different curves correspond to different N/D ratios (lowest: 0.01; highest: 10).

to 10). We see the stability is constant for a fixed N/D ratio, except for some variations in the lowest dimension values for $N/D \geq 5$, caused by random variations in the problem difficulty (very few selected variables on those specific points). For small-sample problem, where $N/D \ll 1$, the stability depends only on the ratio N/D. In real microarray data, the N/D ratio typically ranges from 0.001 to 0.1, which leads to a maximum stability of 0.2, in our simulations. Note that those simulations are based on Gaussian, uncorrelated features, which is one of the easiest classification problems. Moreover we use a selection based on the t-test score, which is the optimal feature selection method in this context. In microarray data, the distribution of the classes is much more complex than Gaussian and the optimal feature selection is unknown. So, the stability on real data should be lower than the stability on artificial data: the values reported on figure 4 should be considered as upper bounds.

B. Real data

We made a set of experiments on three microarray datasets in order to confirm the results obtained on artificial data based simulations. We present here a summary of these results. The error rate (results not shown) decreases exponentially with the number of samples in the lunger cancer and leukemia datasets (respectively, 13.5% for $N = 20$, 8% for $N = 50$, 5% for $N = 150$ and 9.1% for $N = 20$, 4.4% for $N = 30$, 2.7% for $N = 50$), more linearly in the breast cancer dataset (38.2% for $N = 20$, 37.4% for $N = 100$, 36.3% for $N = 200$).

Table I presents stability measures and error rates observed on real datasets for training set sizes of 50 and 100 (or 20 and 35 for the leukemia dataset). The stability measures presented here were computed on strictly non-overlapping sets (pairs of training and test sets) so are not biased by common examples between the different runs. Although stability measures differ in absolute value, they share a same trend, opposite to the error rate. Stability increases with training set size. But globally it remains rather low, although higher than on our artificial data with similar dimensions. This is particularly intriguing in the case of the breast cancer dataset, which has a higher stability but a similar error rate compared to our artificial dataset. These results also confirm the impact of the training set size on the stability of feature selection.

We see that the values of the stability CW_{rel} are higher than the values reported in table 4 for the same N/D ratio. It can be explained by the differences in the experimental design between artificial and real data. In artificial data, the stability is computed from a set of 100 independent datasets. In real data, we have only one datasets that is split in two subsets, this process is repeated 100 times. The 100 splits are not independent, there is therefore a bias that artificially increases the measured stability. Even with this bias, stability values are low.

V. Discussion and Conclusion

In this paper, we have analyzed the performance of gene selection and especially its stability on microarray data. We used existing measures of stability and we introduced ATI_{PA}, a modification of the ATI stability measure adjusted to avoid a bias on the number of selected features. We extensively studied the relation between selection stability and dataset characteristics from a large set of artificial data experiments. We investigated the changes in stability when varying the number of examples, features, selected features and distribution of the discrimination power of features. We show that in small sample problems the probability to select the best features is very low even with the optimal feature selection method. We show empirically that for Gaussian data, the stability depends on the N/D ratio. Since the stability of gene selection from real data is lower than the selection stability on the simple Gaussian data context, we can provide upper bound of stability for real data in function of their size (N and D). The results show that the stability is dramatically low (almost 0) for

TABLE I
CLASSIFICATION ERROR RATE AND SELECTION STABILITY ON THE BREAST CANCER (VAN DE VIJVER), LUNG CANCER (BHATTACHARJEE) AND LEUKEMIA (GOLUB) DATASETS

Measure	**Breast cancer**		**Lung cancer**		**Leukemia**	
	N=50 N/D=0.025	**N=100 N/D=0.05**	**N=50 N/D=0.025**	**N=100 N/D=0.05**	**N=20 N/D=0.003**	**N=35 N/D=0.005**
Error rate	38.2%	37.4%	8.0%	6.1%	9.2%	4.0%
CW_{rel}	0.20	0.26	0.46	0.51	0.24	0.30
ATI_{PA}	0.06	0.10	0.26	0.30	0.13	0.18
$\overline{S_R}$	0.09	0.14	0.51	0.58	0.22	0.26
$\overline{S_W}$	0.33	0.41	0.81	0.85	0.55	0.60

$N/D \leq 0.01$. This leads to the conclusion that for any current microarray data, it is not possible to obtain a stable gene selection for a classification task. These results are coherent with the literature suggesting that thousands of examples are needed to obtain a stable feature selection on microarray data [6], [17]. While this paper focuses on microarray data, it should apply to all genomic data based classification problems.

The conclusions of this work have strong consequences on the development of genomic data based classifiers since we show the classification results are not reproducible. To improve the stability of the gene selection, the first option is simply to increase the number of examples in the datasets. While research projects are necessarily limited in that respect, it seems hopeless to try to construct a stable classifier based only on a few tens of examples. A second way would be to find reliable methods to reduce the dimensionality of the data prior to applying usual filters. For instance, a priori biological knowledge and unsupervised methods could be used to filter some of the irrelevant genes. It could be also interesting to exploit the redundancy among genes. Finally, we point out that the stability measures actually estimate the ability of gene selection methods to produce the same selection of genes. Some studies have been able to successfully reuse the gene selection identified from a microarray dataset on another one, even though this selection was unstable [18]. So, it could be interesting to use an "exportability" measure that estimates if a good gene selection identified on a given dataset remains good on other gene datasets related to the same classification task.

REFERENCES

[1] L. Shi et al, "The MicroArray Quality Control (MAQC)-II study of common practices for the development and validation of microarray-based predictive models," *Nat. Biotechnol.*, vol. 28, pp. 827–838, Aug 2010.

[2] H. S. Chon and J. M. Lancaster, "Microarray-based gene expression studies in ovarian cancer," *Cancer Control*, vol. 18, no. 1, pp. 8–15, Jan 2011.

[3] A. K. Jain and B. Chandrasekaran, "39 dimensionality and sample size considerations in pattern recognition practice," *Handbook of Statistics*, vol. 2, pp. 835–855, 1982.

[4] P. Pudil and P. Somol, "Identifying the most informative variables for decision-making problems - a survey of recent approaches and accompanying problems," *Acta Oeconomica Pragensia*, vol. 2008, no. 4, pp. 37–55, 2008.

[5] J. C. Miecznikowski, D. Wang, S. Liu, L. Sucheston, and D. Gold, "Comparative survival analysis of breast cancer microarray studies identifies important prognostic genetic pathways," *BMC Cancer*, vol. 10, p. 573, 2010.

[6] L. Ein-Dor, O. Zuk, and E. Domany, "Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer," *Proceedings of the National Academy of Sciences*, vol. 103, no. 15, pp. 5923–5928, 2006.

[7] A.-C. Haury and J.-P. Vert, "On the stability and interpretability of prognosis signatures in breast cancer," in *MLSB*. Sašo Džeroski, Simon Rogers and Guido Sanguinetti (Eds.), 10 2010, pp. 27–30.

[8] S. Loscalzo, L. Yu, and C. Ding, "Consensus group stable feature selection," in *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, ser. KDD '09. New York, NY, USA: ACM, 2009, pp. 567–576.

[9] R. D. Shah and R. J. Samworth, "Variable selection with error control: Another look at Stability Selection," *ArXiv e-prints*, May 2011.

[10] P. Somol and J. Novovičová, "Evaluating stability and comparing output of feature selectors that optimize feature subset cardinality," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 11, pp. 1921–1939, 2010.

[11] L. I. Kuncheva, "A stability index for feature selection," in *Artificial Intelligence and Applications*, V. Devedzic, Ed. IASTED/ACTA Press, 2007, pp. 421–427.

[12] P. Krížek, J. Kittler, and V. Hlavác, "Improving stability of feature selection methods," in *Computer Analysis of Images and Patterns*, ser. Lecture Notes in Computer Science, W. Kropatsch, M. Kampel, and A. Hanbury, Eds. Springer Berlin / Heidelberg, 2007, vol. 4673, pp. 929–936.

[13] A. Kalousis, J. Prados, and M. Hilario, "Stability of feature selection algorithms," in *ICDM*. IEEE Computer Society, 2005, pp. 218–225.

[14] A. Bhattacharjee, W. G. Richards, J. Staunton, C. Li, S. Monti, P. Vasa, C. Ladd, J. Beheshti, R. Bueno, M. Gillette, and et al., "Classification of human lung carcinomas by mrna expression profiling reveals distinct adenocarcinoma subclasses." *Proceedings of the National Academy of Sciences of the United States of America*, vol. 98, no. 24, pp. 13 790–13 795, 2001.

[15] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, and C. D. Bloomfield, "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring," *Science*, vol. 286, no. 5439, pp. 531–537, Oct 1999.

[16] M. J. van de Vijver, Y. D. He, L. J. van 't Veer, H. Dai, A. A. Hart, D. W. Voskuil, G. J. Schreiber, J. L. Peterse, C. Roberts, M. J. Marton, M. Parrish, D. Atsma, A. Witteveen, A. Glas, L. Delahaye, T. van der Velde, H. Bartelink, S. Rodenhuis, E. T. Rutgers, S. H. Friend, and R. Bernards, "A gene-expression signature as a predictor of survival in breast cancer," *New England Journal of Medicine*, vol. 347, no. 25, pp. 1999–2009, 2002.

[17] A.-C. Haury, P. Gestraud, and J.-P. Vert, "The influence of feature selection methods on accuracy, stability and interpretability of molecular signatures," *PLoS ONE*, vol. 6, no. 12, p. e28210, 12 2011.

[18] D. Pils, G. Hager, D. Tong, S. Aust, G. Heinze, M. Kohl, E. Schuster, A. Wolf, J. Sehouli, I. Braicu, I. Vergote, I. Cadron, S. Mahner, G. Hofstetter, P. Speiser, and R. Zeillinger, "Validating the impact of a molecular subtype in ovarian cancer on outcomes: A study of the OVCAD Consortium," *Cancer Sci*, Apr 2012.