Classification Complexity in Gene Marker Discovery: analysis on cancer gene expression data

Ivan G. Costa*[1], Thais G. do Rego [1] , Clerton R. A. Filho[1] , Eduardo G. Gusmao[1] and Ana C. Lorena[2] and Marcilio C. P. de Souto[3]

[1] Center of Informatics, Federal University of Pernambuco, Brazil
[2] Center of Mathematics, Computation and Cognition, ABC Fed. Univ., SP, Brazil
[3] Dept. of Informatics and Applied Mathematics, Fed. Univ. of Rio Grande do Norte, Brazil

Email: IGC*- igcf@cin.ufpe.br; TGR - tgr2@cin.ufpe.br; CRAF - craf@cin.ufpe.br; EGG - egg@cin.ufpe.br; ACL - ana.lorena@ufabc.edu.br; MCPS - marcilio@dimap.ufrn.br;

*Corresponding author

Abstract

Background: Classification methods are the main methodology for performing diagnosis and gene marker selection from expression profiles of cancer patients. Nevertheless, gene expression data pose several challenges for these methods. For example, there are high numbers of genes, few patient samples and, in addition, expression values can have noise.

Results: We investigate th suitability of certain statistical indices that measure aspects such as data geometry, topology and complexity of classification boundary to indicate the difficulty in performing classification and gene marker selection in a particular cancer diagnosis data sets. In order to do so, we study the correlation of these indices to the performance of several classification/gene marker selection methods such as SVM, shrunken centroids, discriminant analysis and KNN for several cancer gene expression data sets.

Conclusions: These analysis allow us to investigate the relation of some of the complexity indices in explaining the difficulty of performing cancer diagnosis and marker selection. Therefore, it helps the understanding how the characteristics of these data sets influence the performance of common classification and gene selection methods employed in the literature. Furthermore, some of the indices have a potential to support the selection of a diagnostic/marker methodology for a particular data set.

Background

The measurement of genome-wide expression of patients allows clinical diagnosis to be performed in a molecular level [1, 2]. Given a set of patient expression profiles and its diagnosis, we can use classification methods to find models able to diagnose a new oncoming patient based only on his/her expression profile [1]. Nevertheless, characteristics of gene expression data often makes the task of diagnosis hard. These data sets usually have few patients and measure the expression values of thousands of genes, making the classification problem to lie in a high-dimensional and sparse space. Furthermore, gene expression data have missing values, suffer from several sources of noise and includes patient variability [3].

One important aspect of classification methods is the selection of a subset of genes, whose expression are mostly discriminative for performing the diagnosis of a particular disease. These gene markers could give insights on the molecular mechanisms of the disease. Furthermore, they can be used with small scale profiling technologies, thus enabling the creation of cheap, reliable and easy to perform diagnostic tests [2]. From a methodological point of view, gene selection (or feature selection) could improve the classification error by drastically reducing the gene expression data space. Therefore, data sparsity is lowered and classifiers could have a better generalization in classifying new patients [4].

We have previously performed a study on cancer gene expression data in which we measured statistics of data geometry, topology and shape of the classification boundary to indicate the complexity of the classification task for a given data set [5]. In order to do so, we used a selection of seven classification complexity indices proposed in [6] for general classification problems (see Figure 1 for an example of these complexity statistics). Then, we performed classification using four classical methods considering 10 different gene expression data sets. We found that two complexity indices, Mixture Identifiability and Dimensionality/Sample ratio, were correlated with the test error rates of all classification methods analyzed. In particular, the correlation with the Dimensionality/Sample ratio indicated that the sparsity of the data space is a major aspect in indicating complexity of gene expression-based classification.

In the current work, we extend our previous study by, among other things, performing gene marker selection prior to cancer classification. We employed two gene selection methods, which have been already used in gene expression analysis: sum of squared distances between classes divided by the sum of squares distances within classes (BSS/WSS) and Fisher ratio [7]; and four popular classification methods: Support Vector Machine (SVM) [8], Diagonal Discriminant Analysis (DDA) [9], k-nearest neighbor (KNN) [10] and Shrunken centroids [11]. Furthermore, we use additional data sets to increase the statistical power of our analysis.

We aim to investigate the relation of the complexity indices in explaining the difficulty in performing cancer diagnosis together with gene marker selection. By doing so, we intend to gain insights on the performance of these classification and gene selection techniques, when they are presented to data sets with particular complexity characteristics. Furthermore, we would like to explore, which indices have a potential to support the selection of a diagnostic/gene selection methodology for a particular data set.

Results and Discussion

Comparative Analysis

We performed classification using Shrunken Centroids, SVM, KNN and DDA classifiers with the two feature selection methodologies - SSW/SSB and Fisher Ratio - for the gene expression data bases described in Table 1. Given that SVMs and most of the complexity indices are only able to be employed in classification problems with two classes, we performed one-against-all decompositions for each data set, in which for a problem with C classes, C subproblems are generated, each one distinguishing one class from the remaining classes. This resulted in 85 decomposed data sets. The use of such a decomposition strategy has been indicated in studies analyzing the problem of multi-class classification with SVMs as often leading to overall best classification of cancer gene expression data [12].

For each combination of classification method, gene selection technique and data set decomposition, we performed leave-one-out cross-validation, and measured test error. Feature selection was then performed by a nested cross-validation, selecting for each training data set the top 500, 250, 100, 50, 25 and 10 genes. The Shrunken Centroids method, which automatically performs feature selection, used only data without feature selection. In the end, we kept the best test error and the number of selected genes of a classification method on a data set. These results can be seen in Additional Material 1. More details of the feature selection procedures, classification methods and complexity indices employed are described in the Section "Methods".

Classification Error

The performance of the classifiers can be seen in the Additional Material 1. Overall, Shrunken Centroids and SVM with/without feature selection obtained the lowest classification test errors for most of the data sets: respectively, 33 and 40 out of 85 data sets. In relation to the use of feature selection, 18 of the lowest classification errors were obtained when no feature selection was performed; 20 used the BSS/WSS technique; 7 the Fisher ratio technique and 40 with Shrunken Centroids. These results indicate that in the majority of cases (67 out of 85), classification error was lowered after feature selection.

Classification Complexity

Our main concern here is to compare the error rate of these classifiers generated with the complexity indices describing the data. These results are summarized in Figure 2. This figure presents scatter plots of the complexity indices (y-axis) versus classification error (x-axis) for classification and feature selection methods (see Section "Methods" for a description of the complexity indices). At each plot, points represent the 85 one-against-all decompositions of the data sets. For each method, `1` stands for no feature selection, `2` for BSS/WSS and `3` for Fisher ratio. Numbers in the left upper corner of each scatter plot correspond to the correlation between the error rate and complexity indices. Red values indicate correlations that are statistically significant (t-test with p-value < 0.01).

Two complexity indices measuring linear separability (L1 and L2) have values equal or very close to zero for all data sets (data not shown). This indicates that the high dimensionality makes data to be linearly separable in all "one-against-all" decompositions. Another index that had little discriminative power indicating classification complexity is F2, which measures class overlap. The values of this index were low for most of the data sets, indicating that there is little overlap between classes. The low overlap is likely a result of the data sparsity induced by the high dimensionality of the data sets. We would like to stress that we modified the definition of F2, as it contained a conceptual error in the original definition [6]. See Section Methods for details.

The indices N1, N2 and N3 displayed the most significant correlation with all methods and feature selection techniques with the exception of DDA with Fisher ratio. The positive correlation values evidence that higher values of these indices lead to higher error rates, which indicate more difficult classification problems. Therefore, these indices, which are based on very simple distance based measures, do indeed capture most of the information regarding classification complexity. Still regarding the indices N1, N2 and N3, we can observe that SVM and DDA with the BSS/WSS feature selection presented higher correlations than the other cases. This indicates that the error rate of this feature selection method is more correlated with the data set complexity described by these particular indices.

Another relevant index is $log(D)$, which captures the dimensionality of the original data set previously from feature selection. In this case, the relation with the classification error rate was inverse, that is, high dimensions usually implicated in lower classification error rates. We estimated a similar index $log(D')$, where D' is the number of genes indicated by the feature selection procedure. In this scenario (data not shown), no correlation was found. This indicates that a higher number of genes in the original data set leads classifiers to obtain lower classification errors. On the other hand, we found no relation between classification complexity and the number of genes after feature selection was performed.

Lower F1 values indicate more complex classification problems. The plots show an increase in the error rates with lower F1 values. However, this index did not present a significantly high correlation to many methods. Furthermore, only KNN and SVM with/without feature selection showed positive correlation with the T1 index, which indicates that the dimensionality/number of sample ratios are critical to these classifiers, but not to DDA and Shrunken centroid.

To check if we could combine these indices in a linear manner, we applied principal component analysis to the indices, and included the 1st component as a complexity index. This component represents 67% of the variance in the data. As seen in Figure 2, this index had a high correlation with all methods, except from DDA and Fisher ratio. The correlation of the 1st PCA was similar to the correlation of N1, N2 and N3.

The results presented here are compatible with the ones from our previous study [5]. There N1 and T1 displayed significant correlation with KNN, SVM and other methods. Similar results were observed in the current work, but because of the increase in the number of data sets and complexity indices, we detected here significant correlation of these methods with other complexity indices as well.

Conclusions

The experimental results obtained in this analysis lead us to the following conclusions. The classification complexity indices N1, N2 and N3 successfully explained classification complexity regardless of the method or feature selection used. Another relevant complexity index is the log of the number of genes in the original data set. This indicates that all methods, except DDA, profit from a higher number of genes to perform accurate classification. This suggests that studies of microarray based diagnosis do profit from a larger coverage of genes included in the microarrays. Furthermore, most methods are sensible for the index T1, which measures the ratio between number of genes/number of samples. Thus, a low number of patients and the high dimension of the data sets is another factor limiting the accuracy of these classification methods.

Some findings are in accordance with common knowledge from the gene expression literature, but not yet studied in such large scale expression compendia. First, all data sets decompositions are linearly decomposable, as previously observed in [13]. In relation to the use feature selection, we could see an overall improvement of error rates in 83% of the data sets, as previously observed before [4, 11, 12].

However, we found no relation between the complexity indices, number of genes selected or error rates. This indicates that there is no direct relationship between the error improvement after feature selection and the classification complexity indices used in this study.

Methods

Data

The data of microarray experiments used in this work were collected in GEO (*Gene Expression Omnibus*) [14] or in existing literature. These data sets, which are based on either cDNA or Affymetrix platforms, are a selection of data sets from [15]. The final list of data sets, along with their number of patients, type of microarray platform, class distributions and number of genes are described in Table 1.
We introduce here the basic notation used to describe the cancer gene expression data. Let X be a n by d matrix representing a gene expression data set, where x_{ij} denotes the expression value of sample i and feature (gene) j, x_i is a d-dimensional vector with the expression values of sample (patient) i and $x_{.j}$ is a n-dimensional vector representing the expression values of feature (gene) j. Moreover, Y is an n-dimensional vector, where $y_i \in \{1, .., C\}$ corresponds to the class (or cancer type) of patient i and C is the total number of classes (cancer types) in the data set.
As previously discussed, for data sets with more than 2 classes, we performed one-against all decomposition. That is, for a particular class k, we created a new class vector Y^k, where $y_i^k = 1$ if sample i belongs to class k in the original classification ($y_i = k$) and $y_i^k = 0$ otherwise ($y_i \neq k$). Thus, we obtain C decompositions of data sets, where C is larger than 2.

Pre-processing and Unsupervised Filtering

We apply an unsupervised filter to discard missing values and genes displaying no differential expression within the patients. This filtering, which works without the class information, was applied as in [15]. The pre-processing performed on data from experiments based on the Affymetrix platform has the following steps: (1) values below 10 and above 16000 have been replaced by these bounds, (2) we measure the mean expression of $x_{.j}$, which is represented by $\overline{x}_{.j}$ and eliminate 10% of the highest and lowest values to avoid extreme values (3) each x_{ij} is replaced by $x\prime_{ij} = \log_2(x_{ij}/\overline{x}_{.j})$. In cDNA platform data, it was not necessary to apply transformations, as they were already in logarithmic scale. However, they required the handling of missing values. We eliminated genes that showed more than 10% of missing values. The remaining missing values were substituted by the average expression values of the genes through all tissues.

The unsupervised filter process was as follows: two l and c thresholds were chosen, where the absolute value of the feature $x_{.j}$ has to be higher than l in at least c patients. Genes that do not obey this restriction were excluded from the data set. Column "d" in Table 1 shows the numbers of genes remaining in the data sets after the employment of the former pre-processing and filtering steps.

Supervised Filter

In marker selection, one of the main aspects is the use of statistics to select features (or genes) which best discriminate the classes of patients. We inspect in this work two simple feature selection procedures, which have been successfully applied in previous gene expression studies [7].

The first measure is the sum of squares of squared distances between groups divided by the sum of squares within groups. For feature (gene) j, this measure is:

$$BSS/WSS(j) = \frac{\sum_i^n \sum_k^C I(y_i = k)(\overline{x}_{.j}^k - \overline{x}_{.j})^2}{\sum_i^n \sum_k^C I(y_i = k)(x_{ij} - \overline{x}_{.j}^k)^2} \tag{1}$$

where $\overline{x}_{.j}$ represents the average level of expression of gene j among all samples and $\overline{x}_{.j}^k$ represents the average level of expression of gene j among the samples belonging to class k.

We also use the filter using the Fisher ratio (FR), a measure for (linear) discriminative power of some variable, which is defined as:

$$FR = \frac{\overline{x}_{.j}^{(1)} - \overline{x}_{.j}^{(2)}}{\sigma_j^{2(1)} + \sigma_j^{2(2)}} \tag{2}$$

where $\overline{x}_{.j}^{(k)}$ is the mean of feature j in class k and $\sigma_j^{2(k)}$ is the variance of the class k for feature j [16]. For both feature selection indices, we select the p genes with highest values from a particular training data set to perform classification experiments [7].

Classification Methods

Four classification methods were used for supervised classification in our study: K-nearest neighbors (KNN), support vector machines (SVM), diagonal discriminant analysis (DDA) and nearest shrunken centroid (NSC).

K-nearest neighbors (KNN) algorithm is a method that classifies samples based on the classes of their k nearest samples in the training data [10]. Here, we used the Euclidean distance to find the closest neighbors and set the parameter $k = 3$.

Support Vector Machines (SVMs) is a classification method based on statistical learning theory [8]. Given a dataset, SVMs finds a hyperplane dividing the samples from two classes while also maximizing the distance of these samples to the hyperplane. In particular, we use a Sequential Minimal Optimization (SMO) method for inducing the SVMs implemented in Weka [17]. As indicated in previous studies, most gene expression cancer data sets are linearly separable [13], thus we use simple linear SVMs in our experiments. All other parameters are set as default, as given in the software Weka. It should be also noticed that the SVMs original formulation tackles with binary classification problems. An one-against-all decomposition was employed for multi-class classification.

Diagonal discriminant analysis (DDA) method is based on the use of Gaussians distributions as discriminant rule for classification. Here, we used Gaussians with diagonal class covariance matrices, which assumes independence between features [7]. This method requires no selection of parameters. We used for this method an implementation in the R SMA (Statistics for Microarray Analysis) package.

Nearest Shrunken centroids is a method for centroid based classification, which performs automatic feature selection [11]. The feature selection is based on shrinking centroids of particular features that have a low discrimination power between the classes. The shrinkage factor Δ controls the stringency of the shrinkage factor and set some of the class centroids of particular genes equally, thus discarding these genes from the classifier. For these experiments, we used a nested cross-validation procedure to select the parameter Δ, as implemented in the PAMR (Prediction Analysis for Microarrays in R) package.

In all these methodologies, the final classifier test error was measured with a leave-one-out procedure. Moreover, feature selection was performed in the training data only, to avoid the introduction of bias in the results.

Classification Complexity Measures

We used a selection of indices proposed in [6] and used in [5] to measure classification complexity. Most of these indices are by design only able for accessing data sets with two classes. Therefore, we measure then using a one against all decomposition of the data sets.

We also include in this study a novel complexity index that measures the number of dimensions, which contains most variance in the data. Moreover, we also propose a modification of the Volume of Overlapping Region index proposed in [6], as we found an conceptual flaw on it. We briefly describe all indices below, for more details please refer to [5, 6].

Fisher's Discrimination Ratio ($F1$) calculates the fisher statistics of the features with maximum discriminative in a data set, that is

$$F1 = \max_{j} \frac{\overline{x}_{.j}^{(1)} - \overline{x}_{.j}^{(2)}}{\sigma_j^{2(1)} + \sigma_j^{2(2)}} \tag{3}$$

where $\overline{x}_{.j}^{(k)}$ is the mean of feature j in class k and $\sigma_j^{2(k)}$ is the variance of the class k for feature j. F1 has positive values, where higher values indicates simpler classification problems.

Volume of Overlapping Region ($F2$) measures the length of the overlap between the distributions of values of the classes. For each feature, we measure the area of overlap between classes and normalize it by the total length of the distribution of both classes. Let $max_k(x_{.j})$ and $min_k(x_{.j})$ be the maximum and minimum value of feature j at class k.

$$F2 = \sum_{i}^{N} \frac{\max(0, min(max_1(x_{.j}), max_2(x_{.j})) - max(min_1(x_{.j}), min_2(x_{.j})))}{max(max_1(x_{.j}), max_2(x_{.j})) - min(min_1(x_{.j}), min_2(x_{.j}))} \tag{4}$$

The original definition of the index did not included the leftmost max operation in the nominator. Therefore, the summation in the nominator could have positive terms, when there is overlap within the dimension i, and negative terms when there is a separation in dimension i. Therefore, the final sum can be negative indicating non-overlap in the data set, if the sum of the positive terms (overlaping regions) is smaller than the absolute sum of the negative terms (non-overlaping regions). With this new index, we only consider the area of overlap (or positive term), where 0 will indicate no overlap between classes and positive values will indicate overlap, therefore more complex classification data sets.

Linear Separability Indices ($L1$ and $L2$) access if the classes are linearly separable. This is done with the use of a linear programming method for finding the optimal linear classifier [18]. This method is described in details at [5]. For the linear classifier, we calculate $L1$ by summing the distance of samples to the linear boundary and ($L2$) by estimating its classification error. For both measures, higher values indicates more complex data sets.

Mixture Identifiability Index ($N1$**)** The index $N1$ measures if two classes come from distinct distributions. Given the Euclidean distance of samples, we find a minimum spanning tree and count the proportion of edges connecting samples from distinct classes [19]. Also, higher indices values indicate complex data sets.

Nearest Neighbors Indices ($N2$ **and** $N3$**)** We estimate $N2$ by measuring for each sample the distance of a nearest neighbor in the same class and in the other class. Then, the ratio of the average of all inter and intra class distances is calculated. Thus, $N2$ captures how disperse are intra class samples and how close are inter class samples. For $N3$, we use a leave-one-out nearest neighbor classifier and estimate its error rate. The distance used in indices $N1, N2, N3$ was the Euclidean distance.

Dimensionality/Samples Ratio ($T1$) measures the log of the ratio of number of features against number of samples $T1 = \log(d/n)$.

Authors contributions

TR, CA and EG implemented the approach and performed the experiments. IC, AL and MS designed the study and evaluated the results. All authors wrote the manuscript. All authors read and approved the final manuscript.

Acknowledgements

This work has been partially supported by Brazilian research agencies: FAPESP, FACEPE, CNPq and CAPES.

References

1. Spang R: **Diagnostic signatures from microarrays: a bioinformatics concept for personalized medicine**. *BIOSILICO* 2003, **1**(2):64–68, [http://www.sciencedirect.com/science/article/B75GS-4BRJ67W-J/2/2fa50a82fa348085a698a1b42db4f6a0].
2. van't Veer LJ, Bernards R: **Enabling personalized cancer medicine through analysis of gene-expression patterns.** *Nature* 2008, **452**(7187):564–570, [http://dx.doi.org/10.1038/nature06915].
3. Irizarry RA, Warren D, Spencer F, Kim IF, Biswal S, Frank BC, Gabrielson E, Garcia JGN, Geoghegan J, Germino G, Griffin C, Hilmer SC, Hoffman E, Jedlicka AE, Kawasaki E, Martinez-Murillo F, Morsberger L, Lee H, Petersen D, Quackenbush J, Scott A, Wilson M, Yang Y, Ye SQ, Yu W: **Multiple-laboratory comparison of microarray platforms.** *Nat Methods* 2005, **2**(5):345–350, [http://dx.doi.org/10.1038/nmeth756].
4. Song L, Bedo J, Borgwardt KM, Gretton A, Smola A: **Gene selection via the BAHSIC family of algorithms.** *Bioinformatics* 2007, **23**(13):i490–i498, [http://dx.doi.org/10.1093/bioinformatics/btm216].
5. Costa IG, Lorena AC, y Peres LRMP, de Souto MCP: **Using Supervised Complexity Measures in the Analysis of Cancer Gene Expression Data Sets.** In *BSB*, *Volume 5676 of* Lecture Notes in Computer Science. Edited by Guimarães KS, Panchenko A, Przytycka TM, Springer 2009:48–59, [http://dblp.uni-trier.de/db/conf/wob/bsb2009.html#CostaLPS09].
6. Ho T, Basu M: **Complexity Measures of Supervised Classification Problems**. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 2002, **24**(3):289–300.
7. Dudoit S, Fridlyand J, Speed TP: **Comparison of Discrimination Methods for the Classification of Tumors Using Gene Expression Data**. *Journal of the American Statistical Association* 2002, **97**(457):77–87, [http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.31.2275].
8. Vapnik VN: *The nature of Statistical learning theory.* New York, USA: Springer-Verlag 1995.
9. Hastie T, Tibshirani R, Friedman J: *The elements of statistical learning: Data mining, inference and prediction.* Springer, New York 2001.
10. Ripley BD: *Pattern recognition and neural networks.* Cambridge, UK: University Press 1996.
11. Tibshirani R, Hastie T, Narasimhan B, Chu G: **Diagnosis of multiple cancer types by shrunken centroids of gene expression**. *PNAS* 2002, **99**(10):6567–6572, [http://dx.doi.org/10.1073/pnas.082099299].

12. Statnikov A, Aliferis CF, Tsamardinos I, Hardin D, Levy S: **A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis.** *Bioinformatics* 2005, **21**(5):631–643, [http://dx.doi.org/10.1093/bioinformatics/bti033].

13. Lorena AC, Costa IG, de Souto MCP: **On the complexity of gene expression classification data sets**. In *Proc. of the 8th International Conference on Hybrid Intelligent Systems*, IEEE Computer Society 2008:825–830.

14. Barrett T, Troup DB, Wilhite SE, Ledoux P, Rudnev D, Evangelista C, Kim IF, Soboleva A, Tomashevsky M, Edgar R: **NCBI GEO: mining tens of millions of expression profiles–database and tools update.** *Nucleic Acids Res* 2007, **35**(Database issue), [http://view.ncbi.nlm.nih.gov/pubmed/17099226].

15. de Souto MCP, Costa IG, de Araujo DSA, Ludermir TB, Schliep A: **Clustering cancer gene expression data: a comparative study.** *BMC Bioinformatics* 2008, **9**, [http://dblp.uni-trier.de/db/journals/bmcbi/bmcbi9.html#SoutoCALS08].

16. Yanxiong Peng WL, Liu Y: **A Hybrid Approach for Biomarker Discovery from Microarray Gene Expression Data for Cancer Classification**. *Cancer Inform* 2002, **2**:301–311.

17. Witten IH, Frank E: *Data Mining: Practical machine learning tools and techniques.* Morgan Kaufmann, 2 edition 2005.

18. Smith F: **Pattern Classifier Design by Linear Programming**. *IEEE Transactions on Computers* 1968, **17**(4):367–372.

19. Friedman H, Rafsky LC: **Multivariate generalization of the Wald-Wolfowitz and Smirnov two-sample tests**. *Ann. Statist.* 1979, **7**:697–717.

20. Armstrong SA, Staunton JE, Silverman LB, Pieters R, den Boer ML, Minden MD, Sallan SE, Lander ES, Golub TR, Korsmeyer SJ: **MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia.** *Nat Genet* 2002, **30**:41–47, [http://dx.doi.org/10.1038/ng765].

21. Bhattacharjee A, Richards WG, Staunton J, Li C, Monti S, Vasa P, Ladd C, Beheshti J, Bueno R, Gillette M, Loda M, Weber G, Mark EJ, Lander ES, Wong W, Johnson BE, Golub TR, Sugarbaker DJ, Meyerson M: **Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses.** *Proc Natl Acad Sci U S A* 2001, **98**(24):13790–13795, [http://dx.doi.org/10.1073/pnas.191502998].

22. Chowdary D, Lathrop J, Skelton J, Curtin K, Briggs T, Zhang Y, Yu J, Wang Y, Mazumder A: **Prognostic gene expression signatures can be measured in tissues collected in RNAlater preservative.** *J Mol Diagn* 2006, **8**:31–39.

23. Dyrskjot L, Thykjaer T, Kruh0ffer M, Jensen JL, Marcussen N, Hamilton-Dutoit S, Wolf H, Orntoft TF: **Identifying distinct classes of bladder carcinoma using microarrays.** *Nat Genet* 2003, **33**:90–96, [http://dx.doi.org/10.1038/ng1061].

24. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, Bloomfield CD, Lander ES: **Molecular classification of cancer: class discovery and class prediction by gene expression monitoring.** *Science* 1999, **286**(5439):531–537.

25. Gordon GJ, Jensen RV, Hsiao LL, Gullans SR, Blumenstock JE, Ramaswamy S, Richards WG, Sugarbaker DJ, Bueno R: **Translation of microarray data into clinically relevant cancer diagnostic tests using gene expression ratios in lung cancer and mesothelioma.** *Cancer Res* 2002, **62**(17):4963–4967.

26. Laiho P, Kokko A, Vanharanta S, Salovaara R, Sammalkorpi H, Jarvinen H, Mecklin JP, Karttunen TJ, Tuppurainen K, Davalos V, Schwartz S, Arango D, Makinen MJ, Aaltonen LA: **Serrated carcinomas form a subclass of colorectal cancer with distinct molecular basis.** *Oncogene* 2007, **26**(2):312–320, [http://dx.doi.org/10.1038/sj.onc.1209778].

27. Nutt CL, Mani DR, Betensky RA, Tamayo P, Cairncross JG, Ladd C, Pohl U, Hartmann C, McLaughlin ME, Batchelor TT, Black PM, von Deimling A, Pomeroy SL, Golub TR, Louis DN: **Gene expression-based classification of malignant gliomas correlates better with survival than histological classification.** *Cancer Res* 2003, **63**(7):1602–1607.

28. Pomeroy SL, Tamayo P, Gaasenbeek M, Sturla LM, Angelo M, McLaughlin ME, Kim JYH, Goumnerova LC, Black PM, Lau C, Allen JC, Zagzag D, Olson JM, Curran T, Wetmore C, Biegel JA, Poggio T, Mukherjee S, Rifkin R, Califano A, Stolovitzky G, Louis DN, Mesirov JP, Lander ES, Golub TR: **Prediction of central nervous system embryonal tumour outcome based on gene expression.** *Nature* 2002, **415**(6870):436–442, [http://dx.doi.org/10.1038/415436a].

29. Ramaswamy S, Tamayo P, Rifkin R, Mukherjee S, Yeang CH, Angelo M, Ladd C, Reich M, Latulippe E, Mesirov JP, Poggio T, Gerald W, Loda M, Lander ES, Golub TR: **Multiclass cancer diagnosis using tumor gene expression signatures.** *Proc Natl Acad Sci U S A* 2001, **98**(26):15149–15154, [http://dx.doi.org/10.1073/pnas.211566398].

30. Shipp MA, Ross KN, Tamayo P, Weng AP, Kutok JL, Aguiar RCT, Gaasenbeek M, Angelo M, Reich M, Pinkus GS, Ray TS, Koval MA, Last KW, Norton A, Lister TA, Mesirov J, Neuberg DS, Lander ES, Aster JC, Golub TR: **Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning.** *Nat Med* 2002, **8**:68–74, [http://dx.doi.org/10.1038/nm0102-68].

31. Singh D, Febbo PG, Ross K, Jackson DG, Manola J, Ladd C, Tamayo P, Renshaw AA, D'Amico AV, Richie JP, Lander ES, Loda M, Kantoff PW, Golub TR, Sellers WR: **Gene expression correlates of clinical prostate cancer behavior.** *Cancer Cell* 2002, **1**(2):203–209.

32. Su AI, Welsh JB, Sapinoso LM, Kern SG, Dimitrov P, Lapp H, Schultz PG, Powell SM, Moskaluk CA, Frierson HF, Hampton GM: **Molecular classification of human carcinomas by use of gene expression signatures.** *Cancer Res* 2001, **61**(20):7388–7393.

33. West M, Blanchette C, Dressman H, Huang E, Ishida S, Spang R, Zuzan H, Olson JA, Marks JR, Nevins JR: **Predicting the clinical status of human breast cancer by using gene expression profiles.** *Proc Natl Acad Sci U S A* 2001, **98**(20):11462–11467, [http://dx.doi.org/10.1073/pnas.201162998].

34. Yeoh EJ, Ross ME, Shurtleff SA, Williams WK, Patel D, Mahfouz R, Behm FG, Raimondi SC, Relling MV, Patel A, Cheng C, Campana D, Wilkins D, Zhou X, Li J, Liu H, Pui CH, Evans WE, Naeve C, Wong L, Downing JR: **Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling.** *Cancer Cell* 2002, **1**(2):133–143.

35. Alizadeh AA, Eisen MB, Davis RE, Ma C, Lossos IS, Rosenwald A, Boldrick JC, Sabet H, Tran T, Yu X, Powell JI, Yang L, Marti GE, Moore T, Hudson J, Lu L, Lewis DB, Tibshirani R, Sherlock G, Chan WC, Greiner TC, Weisenburger DD, Armitage JO, Warnke R, Levy R, Wilson W, Grever MR, Byrd JC, Botstein D, Brown PO, Staudt LM: **Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling.** *Nature* 2000, **403**(6769):503–511, [http://dx.doi.org/10.1038/35000501].

36. Bittner M, Meltzer P, Chen Y, Jiang Y, Seftor E, Hendrix M, Radmacher M, Simon R, Yakhini Z, Ben-Dor A, Sampas N, Dougherty E, Wang E, Marincola F, Gooden C, Lueders J, Glatfelter A, Pollock P, Carpten J, Gillanders E, Leja D, Dietrich K, Beaudry C, Berens M, Alberts D, Sondak V: **Molecular classification of cutaneous malignant melanoma by gene expression profiling.** *Nature* 2000, **406**(6795):536–540, [http://dx.doi.org/10.1038/35020115].

37. Bredel M, Bredel C, Juric D, Harsh GR, Vogel H, Recht LD, Sikic BI: **Functional network analysis reveals extended gliomagenesis pathway maps and three novel MYC-interacting genes in human gliomas.** *Cancer Res* 2005, **65**(19):8679–8689, [http://dx.doi.org/10.1158/0008-5472.CAN-05-1204].

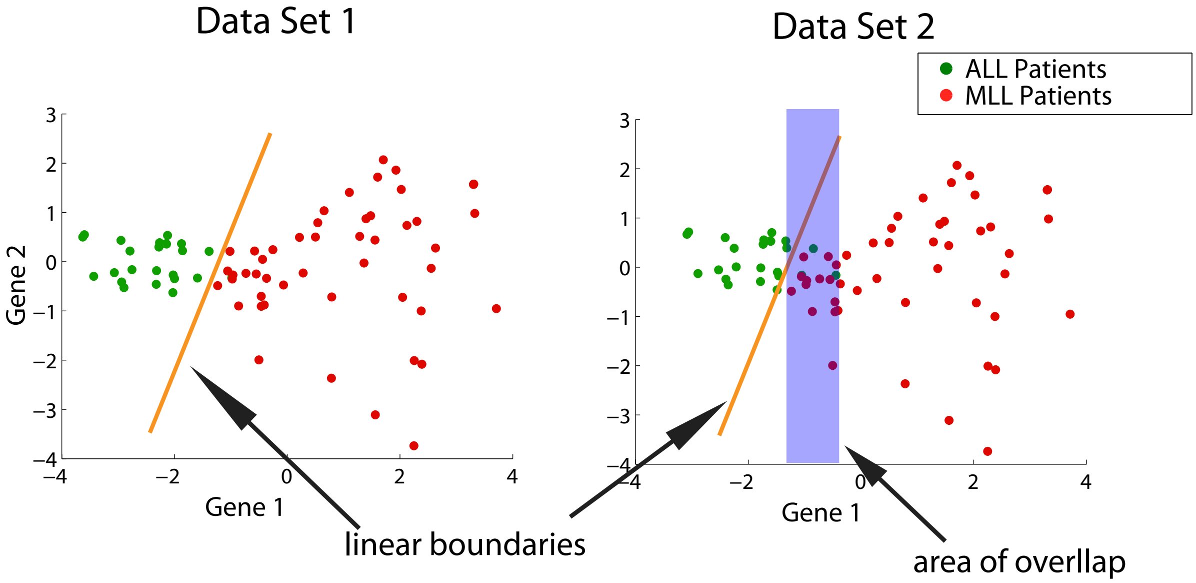
38. Chen X, Cheung ST, So S, Fan ST, Barry C, Higgins J, Lai KM, Ji J, Dudoit S, Ng IO, van de Rijn M, Botstein D, Brown PO: **Gene Expression Patterns in Human Liver Cancers**. *Mol. Biol. Cell* 2002, **13**(6):1929–1939, [http://www.molbiolcell.org/cgi/content/abstract/13/6/1929].

39. Garber ME, Troyanskaya OG, Schluens K, Petersen S, Thaesler Z, Pacyna-Gengelbach M, van de Rijn M, Rosen GD, Perou CM, Whyte RI, Altman RB, Brown PO, Botstein D, Petersen I: **Diversity of gene expression in adenocarcinoma of the lung.** *Proc Natl Acad Sci U S A* 2001, **98**(24):13784–13789, [http://dx.doi.org/10.1073/pnas.241500798].

40. Khan J, Wei JS, Ringner M, Saal LH, Ladanyi M, Westermann F, Berthold F, Schwab M, Antonescu CR, Peterson C, Meltzer PS: **Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks.** *Nat Med* 2001, **7**(6):673–679, [http://dx.doi.org/10.1038/89044].

41. Lapointe J, Li C, Higgins JP, van de Rijn M, Bair E, Montgomery K, Ferrari M, Egevad L, Rayford W, Bergerheim U, Ekman P, DeMarzo AM, Tibshirani R, Botstein D, Brown PO, Brooks JD, Pollack JR: **Gene expression profiling identifies clinically relevant subtypes of prostate cancer.** *Proc Natl Acad Sci U S A* 2004, **101**(3):811–816, [http://dx.doi.org/10.1073/pnas.0304146101].

42. Liang Y, Diehn M, Watson N, Bollen AW, Aldape KD, Nicholas MK, Lamborn KR, Berger MS, Botstein D, Brown PO, Israel MA: **Gene expression profiling reveals molecularly and clinically distinct subtypes of glioblastoma multiforme.** *Proc Natl Acad Sci U S A* 2005, **102**(16):5814–5819, [http://dx.doi.org/10.1073/pnas.0402870102].



43. Risinger JI, Maxwell GL, Chandramouli GVR, Jazaeri A, Aprelikova O, Patterson T, Berchuck A, Barrett JC: **Microarray analysis reveals distinct gene expression profiles among different histologic types of endometrial cancer.** *Cancer Res* 2003, **63**:6–11.

44. Tomlins SA, Mehra R, Rhodes DR, Cao X, Wang L, Dhanasekaran SM, Kalyana-Sundaram S, Wei JT, Rubin MA, Pienta KJ, Shah RB, Chinnaiyan AM: **Integrative molecular concept modeling of prostate cancer progression.** *Nat Genet* 2007, **39**:41–51, [http://dx.doi.org/10.1038/ng1935].

Figures

Figure 1 - Example of complexity classification index

We present here example of complexity indices for two distinct data sets. Each dot represents a patient, the colors (red and green) the two types of cancer and the y and x axis are the expression values of two selected genes. For the data on the right figure, there is an overlap between the classes (blue shaded area), while for the data in the left figure the patients the two classes are clearly separated. Therefore, an index measuring area of overlaping between classes give an indication of the classification complexity, where the higher is the overlap, the more difficult is the classification task. Another way to assess complexity is to estimate a linear boundary which splits the two classes and then measure the error rate of this simple classifier. In the example, we have no error in the left figure, but two green points are wrongly classified in the right figure. Therefore, higher error rates indicate higher complexity.

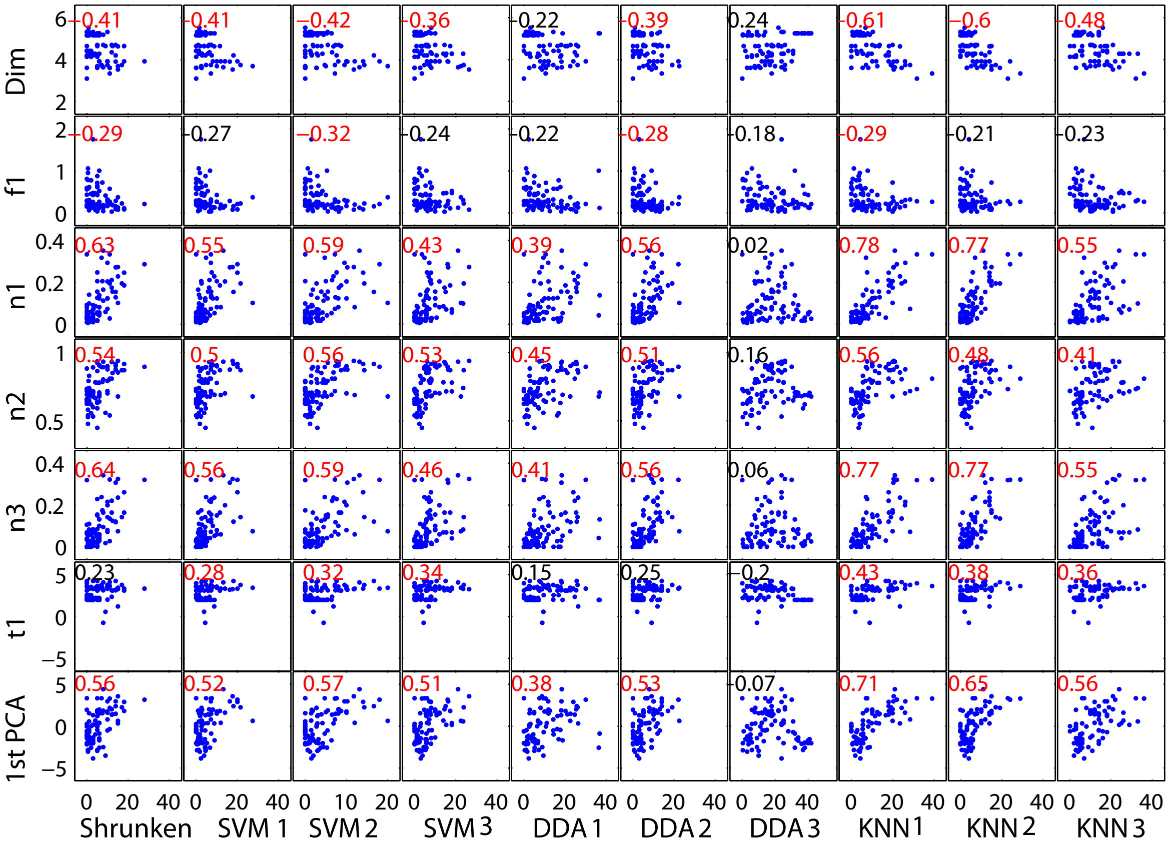


Figure 2 - Correlation between classification methods and complexity indices

Scatter plot of complexity indices versus classification test errors for different combinations of classification and feature selection methods. At each plot, points represent the 85 one-against-all decompositions of the data sets. For each method, 1 stands for no feature selection, 2 for BSS/WSS and 3 for Fisher ratio. Numbers in the left upper corner of each scatter plot correspond to the correlation between the error rate and complexity indices, red values indicates correlations that are statistically significant (t-test with p-value < 0.01).

Dataset	Chip	n	Dist. Classes	d
Armstrong-V2 [20]	Affy	72	24,20,28	2194
Bhattacharjee [21]	Affy	203	139,17,6,21,20	1543
Chowdary [22]	Affy	104	62,42	182
Dyrskjot [23]	Affy	40	9,20,11	1203
Golub-V2 [24]	Affy	72	38,9,25	1877
Gordon [25]	Affy	181	31,150	1626
Laiho [26]	Affy	37	8,29	2202
Nutt-V1 [27]	Affy	50	14,7,14,15	1377
Pomeroy-V1 [28]	Affy	34	25,9	857
Ramaswamy [29]	Affy	190	11,10,11,11,22,10,11,10,30,11,11,11,11,20	1363
Shipp [30]	Affy	77	58,19	798
Singh [31]	Affy	102	58,19	339
Su [32]	Affy	174	26,8,26,23,12,11,7,27,6,28	1571
West [33]	Affy	49	25,24	1198
Yeoh-V1 [34]	Affy	248	43,205	2526
Alizadeh-V1 [35]	cDNA	42	21,21	1095
Alizadeh-V2 [35]	cDNA	62	42,9,11	2093
Bittner [36]	cDNA	38	19, 19	2201
Bredel [37]	cDNA	50	31,14,5	1739
Chen [38]	cDNA	180	104,76	85
Garber [39]	cDNA	66	17,40,4,5	4553
Khan [40]	cDNA	83	29,11,18,25	1069
Lapoint-V2 [41]	cDNA	110	11,39,19,41	2496
Liang [42]	cDNA	37	28,6,3	1411
Risinger [43]	cDNA	42	13,3,19,7	1771
Tomlins-V1 [44]	cDNA	104	27,20,32,13,12	2315

We list the data sets and their characteristics such as type of microarray chip (Chip), number of samples (n), distribution of samples within the classes (Dist. Classes) and dimensionality (*d).*

Tables

Table 1 - Data set description

Additional Files

Additional file 1 — Table with Classification Error, Number of Selected Genes and Complexity Indices

We present here a table with the classification error, number of selected genes and complexity indices against all data set decompositions. The data sets names corresponds to the descriptions in Table 1 and complexity indices and methods to the nomenclature from Figure 2.