Semi-supervised Approach for Finding Cancer Sub-Classes on Gene Expression Data

Clerton Ribeiro[1], Francisco de Assis T. de Carvalho[1], and
Ivan G. Costa[1]

[1]Center of Informatics, Federal University of Pernambuco, Recife, Brazil
craf@cin.ufpe.br, fatc@cin.ufpe.br, igcf@cin.ufpe.br

Abstract. The analysis of cancer gene expression is intrinsically a semi-supervised problem, as one is interested in building a classifier for diagnosis, but also on finding new sub-classes of cancer. We propose here a method for Mixture Discriminant Analysis (MDA), which can simultaneously detect sub-classes of cancer and perform classification. We evaluate the method on 10 gene expression data sets. MDA not only improved the classification in some of these data sets, as it detected some known and putative sub-classes of cancer.

Key words: cancer gene expression, mixture discriminant analysis, semi-supervised learning, constraint based mixture estimation

1 Introduction

The measurement of the expression of all genes of cancer patients has made possible the development of personalized diagnostics [21]. In this context, a standard approach is the use of machine learning methods to build a classifier for a data set with several healthy and cancer patients or with distinct types of cancer [19]. Moreover, analysis on such data sets have shown the presence of unknown sub-types of cancer by the application of clustering methods [1, 10]. Such findings have made the study of gene expression of cancer to be extremely popular, and lead to great advances in cancer diagnosis [21].

These facts indicate that cancer based diagnosis is intrinsically a semi-supervised problem [5]. While the studies generating the gene expression data sets give class labelling of all samples in the data, the frequent discovery of new sub-classes has made the application of both supervised and unsupervised methods routine. Therefore, a method that performs classification of cancer types simultaneously to finding new sub-classes is extremely desirable. By using the detected sub-classes in the classification task, the method can better delineate class boundaries/data distribution, therefore enhancing the overall classification accuracy [11]. Moreover, the detected sub-classes, whenever they are present in the data, are interesting candidates for further analysis by the biomedical experts.

We propose here a semi-supervised method for estimating Mixture Discriminant Analysis (MDA) with Gaussians distributions. MDA, which has been initially proposed in [11], works by fitting a mixture of Gaussian distributions to each class in the data set. One major drawback of this approach is the fact that one needs to estimate the optimal number of components in the mixtures (or sub-classes) for each class independently. This makes the method computationally costly and requires the application of model selection procedures. We propose here the use of a constraint-based-mixture estimation [13] for estimating the MDA. The method has as input the list all negative pairwise constraints, i.e. all pairs of patients that should not be in the same class. The algorithm, which is based on an extension of the Expectation-Maximization (EM) algorithm, searches for solutions with a pre-determined number of groups K satisfying all negative constraints. That is, we do not have patients of distinct classes in a single group, but we allow patients from the same class to belong to several groups. Therefore, if K is higher than the number C of classes (cancer types), the method will return a classifier with $K - C$ novel sub-classes.

A similar approach has been previously shown to work on the classification of time-series of Multiple Sclerosis patients [6]. In this work, we evaluate the MDA method with several data sets from a cancer gene expression compendium [7]. Furthermore, we apply a Quadratic Discriminant Analysis (QDA), which is equivalent to MDA when $K = C$, to serve as a baseline case. To select the optimal number of sub-classes $K - C$, we use a cross-validation procedure. Finally, apply a consensus method proposed in [16] to evaluate if the sub-classes found are stable over distinct solutions obtained by the cross-validation procedure.

2 Material and Methods

2.1 Data Sets

We use in this study 10 public micro-array data sets with cancer gene expression (`http://algorithmics.molgen.mpg.de/Supplements/CompCancer`). An overview of these 10 datasets is presented in Table 1.

Table 1. Data set description

Dataset	**Classes**	n	C	d
`Alizadeh-v2`	DLBCL(42), FL(9), CLL(11)	62	3	4022
`Alizadeh-v3`	DLBCL1(21), DLBCL2(21), FL(9), CLL(11)	62	4	4022
`Armstrong-v1`	ALL(24),MLL(48)	72	2	12582
`Armstrong-v2`	ALL(24), MLL(20), AML(28)	72	3	12582
`Chen`	HCC(104), liver(75)	179	2	22699
`Golub-v1`	ALL(47), AML(25)	72	2	7129
`Golub-v2`	ALL-B(38), ALL-T(9), AML(25)	72	3	7129
`Nutt-v2`	CG(14), NG(14)	28	2	12625
`Nutt-v3`	CO(7), NO(15)	22	2	12625
`Yeoh-v1`	T-ALL(43), B-ALL(205)	248	2	12625

In Table 1, the second column describes the names of the classes (cancer types), as defined in the original publication, and the number of samples (patients) in each class. For further description of classes see [7]. The third column presents the number of samples (n), the fourth column the number of classes and the last column the number of genes (d). It is quite noticeable from the table that all data sets are sparse with a few samples on a high dimensional space.

The data were pre-processed by the application of an unsupervised filter to discard missing values and genes displaying no differential expression, as described in [7]. The pre-processing performed on data from experiments based on the Affymetrix platform (Alizadeh, Golub, Nutt and Yeoh) has the following steps: (1) all values below 10 and above 16000 were replaced by these bounds,(2) we measured the mean expression of each gene and eliminate 10% of the highest and lowest values to avoid extreme values (3) each expression value was replaced by the base 2 log transformation of the ratio between the expression value and the gene mean expression. For cDNA platform data (Armstrong and Chen), it was not necessary to apply transformations, as they were already in logarithmic scale. The unsupervised filter process was as follows: two l and c thresholds were chosen, where the absolute value of the feature has to be higher than l in at least c patients. Genes that do not fit this restriction were excluded from the data set.

2.2 Classification Algorithms

Let X be a d by n matrix representing a gene expression data set, where x_{ij} denotes the expression value of sample (patient) j and feature (gene) i, x_i is a d-dimensional vector with the expression values of sample (patient) i. We also have associated to each data set a vector Y with dimension n, where $y_i \in \{1, ..., C\}$ denotes the class sample i belongs to.

2.3 Discriminant Analysis

Discriminant analysis (DA) methods perform classification by inference over the posterior distribution $\mathbf{P}[y|x]$ [12]. Let $\mathbf{P}[x_i|y_i = c]$ be the class-conditional density modeling the distribution of samples in class c and π_c be the prior distribution of class c, such that $\sum_{c=1}^{C} \pi_c = 1$ and $\pi_c \geq 0$, we can use Bayes Theorem to derive the posterior probability

$$\mathbf{P}[y_i = c|x_i] = \frac{\pi_c \mathbf{P}[x_i|y_i = c]}{\sum_{c'=1}^{C} \pi_{c'} \mathbf{P}[x_i|y_i = c']}. \tag{1}$$

Therefore, classification of a sample x_i can be performed with the rule

$$\hat{y_i} = \underset{c=\{1,...,C\}}{\arg\max}\ \mathbf{P}[y_i = c|x_i]. \tag{2}$$

as given in Eq. 2, where $\hat{y_i}$ is the predicted class for sample i.

The definition of $\mathbf{P}[x_i|y_i = c]$ is application dependent. In gene expression analysis, a usual choice is a multivariate Gaussian density function [9], which is defined as

$$\mathbf{P}[x_i|y_i = c, \theta_c] = \frac{1}{\sqrt{(2\pi)^d |\Sigma_c|}} \exp^{\frac{1}{2}(x_i-\mu_c)^{\mathbf{T}} \Sigma_c^{-1} (x_i-\mu_c)}, \tag{3}$$

where θ_c are the parameters (μ_c, Σ_c). μ_c and Σ_c can be estimated with the mean and covariance matrices of samples of class c and $\pi_c = n_c/n$, where n_c is the number of samples in class c [12].

Given sparsity of the data (few samples and high dimension), it is usual to assume independence among the attributes given the class. In gene expression analysis this is done by estimating a diagonal parameterization of the covariance matrix Σ_c, i.e. only the diagonal entries are estimated and all other values are set to zero [9]. This variant of DA is known as Diagonal Quadratic Discriminant Analysis (DQDA) and will be used in this study as a baseline method.

2.4 Mixture Discriminant Analysis

With mixture of discriminant analysis (MDA), we assume that class condition densities can be defined as a mixture model, that is

$$\mathbf{P}[x_i|y_i = c] = \sum_{k=1}^{K} \alpha_k \mathbf{P}[x_i|z_i = k], \tag{4}$$

where $\alpha_k, i = 1, ..., K$ are the mixing coefficients. In [11], the estimation of these mixture were performed with the application of the EM algorithm for each class to be classified.

2.5 Mixture Model Estimation with Constraints

A standard mixture model can be defined as

$$\mathbf{P}[x_i|\Theta] = \sum_{k=1}^{K} \alpha_k \mathbf{P}[x_i|y_i = k, \theta_k] \tag{5}$$

as given in Eq. 5,where $\Theta = (\alpha_1, ..., \alpha_k, \theta_1, ..., \theta_K)$ are the model parameters and α_k are the mixing coefficients. By including a set of hidden labels represented by the n-dimensional vector Z, where $z_i \in \{1, .., K\}$ defines the component generating the x_i, we obtain the complete data likelihood

$$\mathbf{P}[X, Y|\Theta] = \mathbf{P}[X|Z, \Theta]\mathbf{P}[Z|\Theta]. \tag{6}$$

We can use then the `EM` method to estimate the parameters Θ and component assignments Z maximizing the complete likelihood (see [15] for details).

In constrained-based-mixture estimation (and its similar constrained based clustering), the user can define a $n \times n$ matrix W with negative pairwise constraints, where $w^-_{ij} = 1$ if samples i and j should not belong to the same mixture component and $w^-_{ij} = 0$ otherwise. The constraints are incorporated in the estimation by extending the prior probability of the hidden variable to $\mathbf{P}[Z|\Theta, W] = \mathbf{P}[Z|\Theta]\mathbf{P}[W|Z]$. Assuming $\mathbf{P}[W|Z]$ follows a Gibbs distribution, there is a variation of the EM algorithm for estimating Z and Θ [13, 14]. The method requires the redefinition of the posterior assignment distribution as

$$\mathbf{P}[z_i = k|x_i, W] = \frac{\pi_c \mathbf{P}[x_i|z_i = k]}{\mathcal{Z}} \exp^{\sum_{j \neq i} -\lambda^- w_{ij}^- \mathbf{P}[z_j = k|x_j, W]}, \tag{7}$$

where $\mathcal{Z} = \sum_{k=1}^{K} \mathbf{P}[z_i = k|x_i, W]$ and λ^- is the Lagrange parameter defining the penalty weight of constraints violations.

2.6 Constraint-based Mixture Discriminant Analysis

We propose here the use of the constraint-based mixture estimation method described above for obtaining a MDA classifier. By setting the penalty parameter λ^- with a high value and the constraint matrix W, such that $w_{ij}^- = 1$ if $y_i \neq y_j$ and $w_{ij}^- = 0$ otherwise, we will obtain solutions where samples with distinct classes are not in the same mixture component. Furthermore, by choosing a number of components $K > C$, some of the classes will be related to more than one mixture component. In other words, the mixture will divide some of the classes in sub-classes.

Therefore, we need a procedure to relate the mixture components with the classes. This can be achieved by relating the assignment vector Z of the mixture with the class vector Y. We can estimate the probability of obtaining class c given component k by

$$\mathbf{P}[y = c|z = k] = \frac{\sum_{i=1}^{N} \mathbf{1}(y_i = c)\mathbf{1}(z_i = k)}{\sum_{i=1}^{N} \mathbf{1}(z_i = k)}, \tag{8}$$

where $\mathbf{1}$ is the identity function. From this, we can define the mapping

$$\text{ClassOf}(k) = \underset{c=\{1,\dots,C\}}{\arg\max}\ \mathbf{P}[y = c|z = k], \tag{9}$$

which defines the class c related to component k.

We can use this mapping and parameters Θ, which has been estimated with the method described in Section 2.5, to define the class conditionals as defined in Eq. 4 and obtain a MDA classifier with the use of Eq. 1.

2.7 Experimental Design and Consensus Analysis

For each data set, we performed a leave-one-out cross-validation. All accuracies described in the following are based on the test set alone. Then we use the Friedman test followed by a multiple comparison correction procedure to access the significance of the ranking of the methods [8]. For the final interpretation of the sub-classes, we need a method for combining the results of the classifiers (training and test sets) for all leave one out runs. For this task, we use a procedure proposed in [4, 16]. First, we build a co-occurrence matrix by counting for each pair of samples the number of times they appear in the same component across the different solutions Z. The consensus method works by reshuffling the matrix and clustering samples that share similar groups over solutions [16].

3 Experiments and Results

We investigate here if the use of the Mixture Discriminant Analysis method improves classification accuracy in relation to the baseline method DQDA, which is the equivalent to MDA when $K = C$. Data sets, where the MDA improves or sustains the classification accuracy, are of interest, as these indicate the presence of sub-classes of cancer.

Table 2. Accuracy and standard deviation from classification methods for each data set

Dataset	DQDA	MDA $c+1$	MDA $c+2$
`Alizadeh-v1`	**95.24** (21.55)	80.95 (39.74)	80.95 (26.07)
`Alizadeh-v2`	96.77 (17.81)	**100** (0)	**100** (0)
`Armstrong-v1`	98.61 (11.79)	97.22 (16.55)	98.61 (11.79)
`Armstrong-v2`	**94.44 (23.07)**	**94.44 (23.07)**	88.89 (11.79)
`Chen`	91.62 (27.79)	91.06 (28.61)	94.41 (20.72)
`Golub-v1`	**98.61 (11.79)**	97.22 (16.55)	93.05 (16.55)
`Golub-v2`	90.28 (29.83)	90.27 (29.83)	90.27 (20.12)
`Nutt-v2`	78.57 (41.79)	71.42 (46.00)	82.14 (31.50)
`Nutt-v3`	86.36 (35.13)	90.9 (29.42)	81.81 (38.56)
`Yeoh-v1`	**96.16 (21.50)**	92.74 (26.00)	91.93 (16.60)

We depict the accuracies and standard deviation in Table 1. Values in bold face represent the method, which obtained a statistically significant improvement as indicated by the Friedman test [8]. For three datasets (Alizadeh-v1, Golub-v1 and Yeoh-v1), DQDA obtained best results. In Alizadeth-v2 MDA with c+1 and c+2 obtained better results and in Armstrong-v2 both DQDA and MDA c+1 were best. In all other cases, there was no statistically relevant difference. Note that we used a leave-one-out cross-validation, due of the small number of samples in the data sets. Such setting, usually lead to low accuracy bias but high deviation, lowering the statistical power of comparisons [3].

As expected, MDA did not obtained a higher accuracy than DQDA in all data sets, not all data sets contain sub-classes. Moreover, the limited number of patients may lead to over-fitting with solutions with many sub-classes (too complex models). In some scenarios MDA was better or equivalent to DQDA. As the existence of sub-classes is interesting from the application problem, we prefer the solution of MDA with more components, whenever accuracy is equivalent to DQDA.

Some of the data sets above, Alizadeh-v1, Armstrong-v2 and Gollub-v2, represent the original classification performed by the specialists, which were latter found to contain sub-classes with the use of unsupervised methods [1, 2, 10]. In these scenarios, MDA had superior or equivalent accuracies in relation the QDA.

To assess if MDA is successful in detecting the sub-classes, we perform the consensus analysis [4, 16] on the Armstrong-v2 data set. In Figure 1, we depict the co-occurrence matrix, where a particular entry indicates the number of times the pair

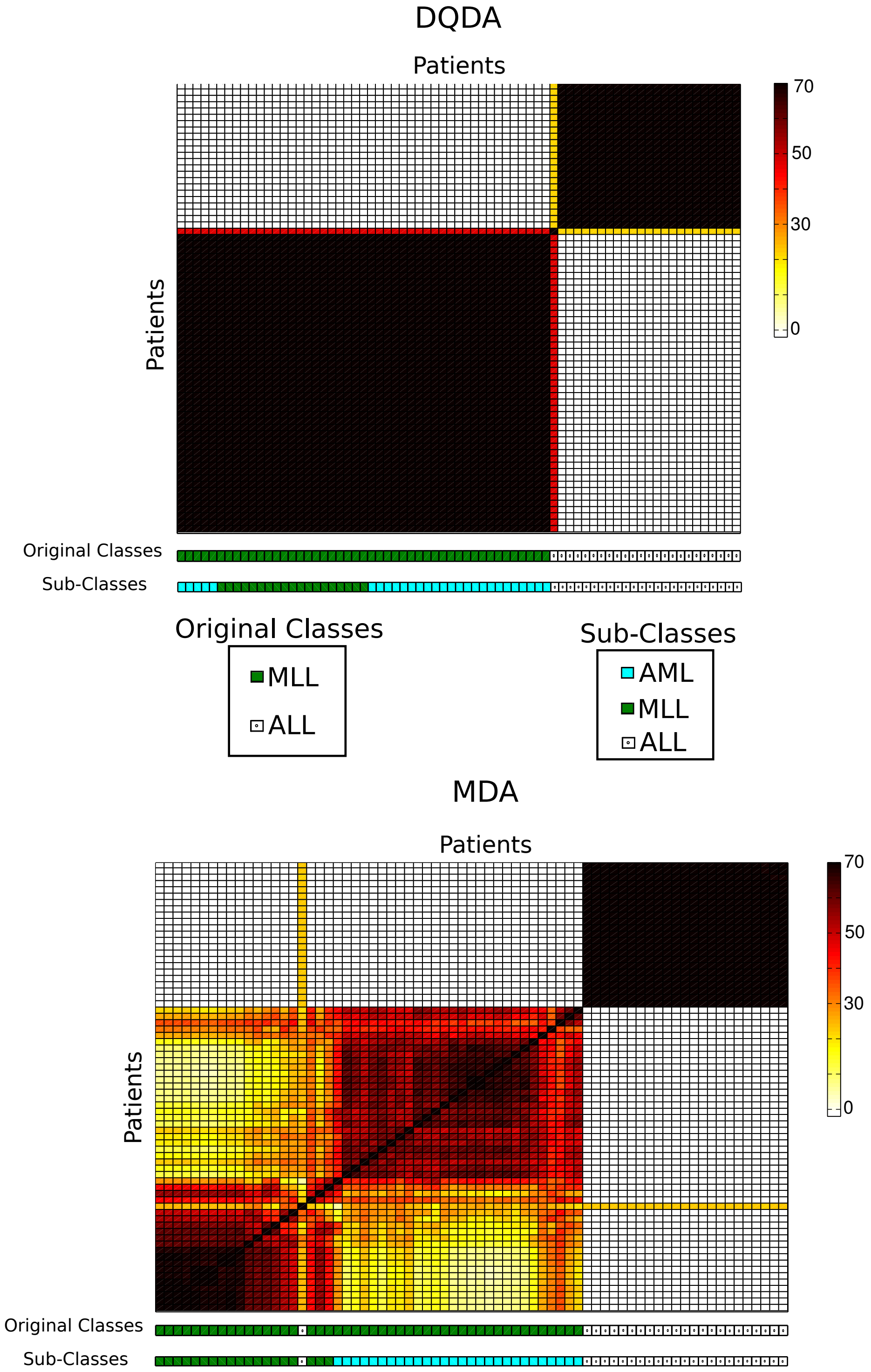

Fig. 1. Consensus Analysis on the Armstrong-v2 data for DQDA (top) and MDA $C+1$ (bottom).

of patients were classified in the same class/sub-class (darker values indicate higher counts). Ideally, the consensus matrix should a block of dark values for each class indicating that the same patients were consistently classified together. As seen in Figure 1 top, DQDA obtained an almost perfect classification and separated all but one patient from the original classes: lymphoblastic leukemias with MLL translocations (MLL) and Acute lymphoblastic leukemias (ALL) [2]. This is indicated in the figure by the two block of dark values.

The original study applied a clustering algorithm and found that 28 patients, which were originally classified as patients with MLL, had distinct expression signatures from other MLL patients [2]. These had their diagnostics changed to akute myelogenous leukemias (AML). As indicated in Figure 1 bottom, MDA with $c + 1$ components, detected the subclasses AML and MLL as indicated by the two blocks of dark values in the left-bottom part of the matrix. Note that in this data set, only the two original classes (MLL and ALL) were given as input for the constraints. This exemplifies a case when MDA successfully finds sub-classes.

Another interesting data set is Nutt-v3, where we see a improvement on the classification accuracy of MDA in relation to DQDA. Moreover, the co-occurrence analysis indicated two sub-classes of patients with non-classic anaplastic oligodendrogliomas, with respectively 11 and 4 patients. These non-classic gliomas are of difficult diagnosis and these sub-classes have not been previously reported in the original study [17]. We detected a significant difference (t-test with p-value < 0.05) in the patient survival time: 672 days for sub-class 1 and 1079 days for sub-class 2.

Next, we explored the genes (features) that are discriminative between these sub-classes by estimating the Fisher discriminant ratio for all genes and ranking them. We selected the 50 most discriminant genes for each class and we performed an enrichment analysis with the g:profiler tool [18]. The analysis revealed that genes up-regulated in sub-class 1 are related to metabolic process and cell cycle, while genes over-expressed in sub-class 2 are related to immune response. These indicate a quite distinct expression signature of these sub-classes, possibly as a result of distinct immune response of the patients to cancer. However, further patient and clinical data are required for the validation of the potential sub-classes.

4 Final Remarks

We propose a new method for estimation of mixture discriminant analysis. This methods improves the original proposal of MDA [11] by requiring only one pass of the EM algorithm to obtain solutions. In the analysis of cancer gene expression, we have shown that MDA can improve classification and successfully indicate the existence of sub-classes of cancer of gene expression data sets. This was exemplified on the classical study from Armstrong et al. [2]. Moreover, interesting sub-classes of non-classical gliomas were found in the Nutt data set.

As future work, we would like to either include new data sets in the study and perform a more detailed biological analysis of the sub-classes found. From a methodological point of view, the MDA can be improved by the use of feature selection methods to cope with the high-dimensionality problem, for example using an approach similar to Shrunken centroids [20].

Acknowledgment

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

References

1. A. A. Alizadeh, M. B. Eisen, R. E. Davis, C. Ma, I. S. Lossos, A. Rosenwald, J. C. Boldrick, H. Sabet, T. Tran, X. Yu, J. I. Powell, L. Yang, G. E. Marti, T. Moore, J. Hudson, L. Lu, D. B. Lewis, R. Tibshirani, G. Sherlock, W. C. Chan, T. C. Greiner, D. D. Weisenburger, J. O. Armitage, R. Warnke, R. Levy, W. Wilson, M. R. Grever, J. C. Byrd, D. Botstein, P. O. Brown, and L. M. Staudt. Distinct types of diffuse large b-cell lymphoma identified by gene expression profiling. *Nature*, 403(6769):503–511, Feb 2000.
2. S. A. Armstrong, J. E. Staunton, L. B. Silverman, R. Pieters, M. L. den Boer, M. D. Minden, S. E. Sallan, E. S. Lander, T. R. Golub, and S. J. Korsmeyer. Mll translocations specify a distinct gene expression profile that distinguishes a unique leukemia. *Nat Genet*, 30(1):41–47, Jan 2002.
3. U. M. Braga-Neto and E. R. Dougherty. Is cross-validation valid for small-sample microarray classification? *Bioinformatics*, 20(3):374–380, Feb 2004.
4. J.-P. Brunet, P. Tamayo, T. R. Golub, and J. P. Mesirov. Metagenes and molecular pattern discovery using matrix factorization. *Proc Natl Acad Sci U S A*, 101(12):4164–4169, Mar 2004.
5. O. Chapelle, B. Schölkopf, and A. Zien, editors. *Semi-Supervised Learning*. MIT Press, Cambridge, MA, 2006.
6. I. G. Costa, A. SchÃ¶nhuth, C. Hafemeister, and A. Schliep. Constrained mixture estimation for analysis and robust classification of clinical time series. *Bioinformatics*, 25(12):i6–14, Jun 2009.
7. M. C. P. de Souto, I. G. Costa, D. S. A. de Araujo, T. B. Ludermir, and A. Schliep. Clustering cancer gene expression data: a comparative study. *BMC Bioinformatics*, 9:497, 2008.
8. J. Demšar. Statistical comparisons of classifiers over multiple data sets. *J. Mach. Learn. Res.*, 7:1–30, 2006.
9. S. Dudoit, J. Fridlyand, and T. P. Speed. Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of the American Statistical Association*, 97(457):77–87, 2002.
10. 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, C. D. Bloomfield, and E. S. Lander. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 286(5439):531–537, Oct 1999.
11. T. Hastie and R. Tibshirani. Discriminant analysis by gaussian mixtures. *Journal of the Royal Statistical Society, Series B*, 58:155–176, 1996.
12. T. Hastie, R. Tibshirani, and J. Friedman. *The elements of statistical learning: Data mining, inference and prediction*. Springer, New York, 2001.
13. T. Lange, M. H. Law, A. K. Jain, and J. M. Buhmann. Learning with constrained and unlabelled data. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 1, pages 731–738, 2005.

14. Z. Lu and T. Leen. Semi-supervised learning with penalized probabilistic clustering. In L. K. Saul, Y. Weiss, and L. Bottou, editors, *Advances in Neural Information Processing Systems 17*, pages 849–856. MIT Press, 2005.
15. G. MacLachlan and D. Peel. *Finite Mixture Models*. Wiley Series in Probability and Statistics. Wiley, 2000.
16. S. Monti, P. Tamayo, J. P. Mesirov, and T. R. Golub. Consensus clustering: A resampling-based method for class discovery and visualization of gene expression microarray data. *Machine Learning*, 52(1-2):91–118, 2003.
17. C. L. Nutt, D. R. Mani, R. A. Betensky, P. Tamayo, J. G. Cairncross, C. Ladd, U. Pohl, C. Hartmann, M. E. McLaughlin, T. T. Batchelor, P. M. Black, A. von Deimling, S. L. Pomeroy, T. R. Golub, and D. N. Louis. Gene expression-based classification of malignant gliomas correlates better with survival than histological classification. *Cancer Res*, 63(7):1602–1607, Apr 2003.
18. J. Reimand, M. Kull, H. Peterson, J. Hansen, and J. Vilo. g:profiler–a web-based toolset for functional profiling of gene lists from large-scale experiments. *Nucleic Acids Res*, 35(Web Server issue):W193–W200, Jul 2007.
19. R. Spang. Diagnostic signatures from microarrays: a bioinformatics concept for personalized medicine. *BIOSILICO*, 1(2):64–68, May 2003.
20. R. Tibshirani, T. Hastie, B. Narasimhan, and G. Chu. Diagnosis of multiple cancer types by shrunken centroids of gene expression. *PNAS*, 99(10):6567–6572, May 2002.
21. L. J. van't Veer and R. Bernards. Enabling personalized cancer medicine through analysis of gene-expression patterns. *Nature*, 452(7187):564–570, Apr 2008.