SETIT 2007
4^{th} International Conference: Sciences of Electronic,
Technologies of Information and Telecommunications
March 25-29, 2007 - TUNISIA

Randomized Embedding Cluster ensembles for gene expression data analysis

Alberto Bertoni *, Giorgio Valentini *

DSI, Dipartimento di Scienze dell' Informazione, Università degli Studi di Milano, Via Comelico 39, 20135 Milano, Italia.

bertoni@dsi.unimi.it
valentini@dsi.unimi.it

Abstract In the framework of unsupervised pattern analysis of gene expression, the high dimensionality of the data as well as the accuracy of clustering algorithms and the reliability of the discovered clusters are critical problems. We propose and analyze an algorithmic scheme for unsupervised cluster ensembles, where the dimensionality reduction is obtained by means of randomized embeddings with low distortion. Multiple "base" clusterings are performed on random subspaces, approximately preserving the distances between the projected examples. In this way the accuracy of each "base" clustering is maintained, and the diversity between them is improved. By combining the multiple clusterings, we can enhance the overall accuracy and the reliability of the discovered clusters, as shown by our experimental results with high-dimensional gene expression data.

Keywords: clustering, DNA microarray analysis, random projections, unsupervised ensembles.

Introduction

Exploratory unsupervised analysis of gene expression data may discover functional classes of tissue specimens at bio-molecular level. Relevant applications include discovery of new subclasses of diseases that may be critical for a more refined bio-molecular diagnosis and for appropriate treatments tailored to the bio-molecular portrait of a patient [Rosenwald et al., 2002]. Clustering algorithms play also a significant role with supervised classification of DNA microarray data (e.g. for the bio-molecular diagnosis of tumors, [Furey et al., 2000]), since the labels of the classes are often determined through unsupervised clustering methods. As a consequence, inaccurate cluster assignments could lead to erroneous diagnoses and inappropriate treatment protocols with a a consequent impact on healthiness and survival of patients [Dudoit and Fridlyand, 2002].

The main goal of this work is to improve the accuracy of clustering algorithms in the framework of gene expression data analysis, through unsupervised ensemble methods. As a by-product of our approach, a set of statistics to evaluate the stability of the obtained clusters are also given.

Ensemble clustering methods have been recently applied to gene expression data analysis, to improve accuracy, robustness and stability of the discovered clusters [Monti et al., 2003]. [Hu

and Yoo, 2004] combined different clustering algorithms to obtain a more stable consensus partition of the data, while [Dudoit and Fridlyand, 2003] applied resampling techniques borrowed from classical supervised bagging techniques to improve the accuracy of clustering algorithms.

Our approach to ensemble clustering exploits one of the characteristics of DNA microarray data that make difficult to process them, that is their high dimensionality. Indeed it is well-known that one of the main problem of gene expression data processing is represented by their high dimensionality and relatively low cardinality: in this context the "curse of dimensionality" problem arises [Bellman, 1961]. The main supervised approach to this problem consists in reducing the dimensionality through gene selection methods (see [Guyon and Elisseeff, 2003] for a recent review). When we need to discover unknown common patterns of expression or new subclasses of diseases (e.g. identifying molecular variations among tumors for a finer and more reliable classification), this approach is not applicable, because we do not known the label of the examples in advance. In this unsupervised context Principal Component Analysis may be in principle applied to reduce dimensionality, but useful discriminant information may be lost. Recently [Smolkin and Gosh, 2003] proposed an approach based on an unsupervised version of the random subspace method [Ho, 1998] to assess the reliability of the discovered gene expression clusters. By extending this approach to more general random projections, in the framework of random embeddings between euclidean spaces, we propose an ensemble method based on multiple clusterings of the data, performed in subspaces of reduced dimension and with low metric distortion.

The next section introduces some basic concepts about randomized embeddings, in particular focusing on low distorted randomized embeddings and random projections. Then the proposed *Randomized embedding clustering (RECLUST)* ensemble algorithm scheme is presented. Sect. 3 show the results of the application of the ensemble method to high dimensional DNA microarray data. The discussion and the conclusions end the paper.

1 Randomized embeddings

1.1 Clustering and data compression

Consider a data set $X = \{x_1, x_2, \ldots, x_n\}$, where $x_i \in \mathbb{R}^d$, $(1 \leq i \leq n)$; a subset $A \subseteq \{1, 2, \ldots, n\}$ univocally individuates a subset of examples $\{x_j | j \in A\} \subseteq X$. The data set X may be represented as a $d \times n$ matrix D, where columns correspond to the examples (e.g. patients), and rows correspond to the "components" of the examples $x \in X$ (e.g. the gene expression levels for different d genes).

A *k-clustering* C of X is a list $C =< A_1, A_2, \ldots, A_k >$, with $A_i \subseteq \{1, 2, \ldots, n\}$ and such that and $\bigcup A_i = \{1, \ldots, n\}$.

A *clustering algorithm* $\mathcal{C}$ is a (possibly randomized) procedure that, having as input a data set X and an integer k, outputs a k-clustering C of X: $\mathcal{C}(X, k) =< A_1, A_2, \ldots, A_k >$. We may also equivalently apply a clustering algorithm to the matrix D that represents X, having that $\mathcal{C}(D, k) = \mathcal{C}(X, k)$. Here we suppose that the result depends only on the the distances $||x_i - x_j||$ between elements in X.

The computation time of a clustering algorithm $\mathcal{C}$ depends critically on the dimension d of the elements in X. In order to compress the data set, we need to find a linear map $\mu : \mathbb{R}^d \to \mathbb{R}^{d'}$, with $d' < d$ such that $||\mu(x_i) - \mu(x_j)|| \simeq ||x_i - x_j||$, for $x_i, x_j \in X$. In this way, algorithms whose results depend only on the distances $||x_i - x_j||$ could be applied to the compressed data $\mu(X)$, giving the same results.

Unfortunately, in general, such embeddings do not exist, but we can obtain the desired result

even if a certain distortion is introduced; to this end, *randomized embeddings with low distortion* represent a key concept.

1.2 Randomized embeddings with low distortion.

For all $x, y \in X$ the *distortion* $dist_\mu(x,y)$ is defined:

$$dist_\mu(x,y) = \frac{||\mu(x) - \mu(y)||}{||x-y||} \qquad (1)$$

A *randomized embedding* between $\mathbb{R}^d$ and $\mathbb{R}^{d'}$ with distortion $1+\epsilon$, $(0 < \epsilon \leq 1/2)$ and failure probability P is a distribution probability on the linear mapping $\mu : \mathbb{R}^d \to \mathbb{R}^{d'}$, such that, for every pair $p, q \in \mathbb{R}^d$, the following property holds with probability $\geq 1 - P$:

$$\frac{1}{1+\epsilon} \leq \frac{||\mu(p) - \mu(q)||}{||p-q||} \leq 1+\epsilon \qquad (2)$$

The main result on randomized embedding is due to [Johnson and Lindenstrauss, 1984], who proved the existence of a randomized embedding $\mu : \mathbb{R}^d \to \mathbb{R}^{d'}$ with distortion $1+\epsilon$ and failure probability $e^{\Omega(-d'\epsilon^2)}$, for every $0 < \epsilon < 1/2$. As a consequence, for a fixed data set $S \subset \mathbb{R}^d$, with $|S| = n$, by union bound, for all $p, q \in S$, it holds:

$$Prob\left(\frac{1}{1+\epsilon} \leq \frac{||\mu(p) - \mu(q)||}{||p-q||} \leq 1+\epsilon\right) \geq 1 - n^2 e^{\Omega(-d'\epsilon^2)} \qquad (3)$$

Hence, by choosing d' such that $n^2 e^{\Omega(-d'\epsilon^2)} < 1/2$, it is proved the following:

Johnson-Lindenstrauss (JL) lemma: Given a set S with $|S| = n$ there exists a $1+\epsilon$-distortion embedding into $\mathbb{R}^{d'}$ with $d' = c \log n/\epsilon^2$, where c is a suitable constant.

The embedding exhibited in [Johnson and Lindenstrauss, 1984] consists in random projections from $\mathbb{R}^d$ into $\mathbb{R}^{d'}$, represented by matrices $d' \times d$ with random orthonormal vectors. Similar results may be obtained by using simpler embeddings [Bingham and Mannila, 2001], represented through random $d' \times d$ matrices $P = 1/\sqrt{d'}(r_{ij})$, where r_{ij} are random variables such that:

$$E[r_{ij}] = 0, \qquad Var[r_{ij}] = 1$$

For sake of simplicity, we call random projections even this kind of embeddings.

1.3 Random projections.

Suppose that $d' = c \log n/\epsilon^2 << d$; the *JL lemma* guarantees the existence of a $d' \times d$ matrix P such that the columns of the "compressed" data set $D_P = PD$ have approximately the same distance (up to a distortion $1+\epsilon$) of the corresponding columns in D. Moreover there is a randomized algorithm that, having in input D, outputs D_P in time $\mathcal{O}(dd'n)$ with high confidence.

Examples of randomized maps are:

1. *Plus-Minus-One (PMO)* random projections: represented by $d' \times d$ matrices $P = 1/\sqrt{d'}(r_{ij})$, where r_{ij} are uniformly chosen in $\{-1, 1\}$, such that $Prob(r_{ij} = 1) = Prob(r_{ij} = -1) = 1/2$. In this case the *JL lemma* holds with $c \simeq 4$.

2. *Achlioptas* random projections [Achlioptas, 2001]: represented by $d' \times d$ matrices $P = 1/\sqrt{d'}(r_{ij})$, where r_{ij} are chosen in $\{-\sqrt{3}, 0, \sqrt{3}\}$, such that $Prob(r_{ij} = 0) = 2/3$, $Prob(r_{ij} = \sqrt{3}) = Prob(r_{ij} = -\sqrt{3}) = 1/6$. In this case also we have $E[r_{ij}] = 0$ and $Var[r_{ij}] = 1$ and the *JL lemma* holds.

3. *Random Subspace (RS)* [Ho, 1998]: represented by $d' \times d$ matrices $P = \sqrt{d/d'}(r_{ij})$, where r_{ij} are uniformly chosen with entries in $\{0, 1\}$, and with exactly one 1 per row and at most one 1 per column. It is worth noting that, in this case, the "compressed" data set $D_P = DX$ can be quickly computed in time $\mathcal{O}(nd')$, independently from d. Unfortunately, *RS* does not satisfy the *JL lemma*.

2 The *RE-Clust* ensemble algorithm

In this section we introduce the cluster ensemble algorithm *RE-Clust* (acronym for Random-

ized Embedding Clustering). It is based on three main items.

1. *Data compression.* Clustering algorithms work on the basis of the dissimilarities or distances between examples, and randomized maps allow to embed data into lower dimensional spaces, approximately preserving their distances.

2. *Multiple "base" clusterings on multiple instances.* For a fixed randomized map, we can obtain multiple instances of data sets by applying multiple random projections to the same input data set. Multiple "base" clusterings are then produced, by calling a clustering algorithm on the obtained multiple instances.

3. *Combining multiple clusterings.* The final clustering is produced by combining the multiple "base" clusterings, In principle, we may combine clusterings using a "majority voting" approach, or adapting other combination schemes previously proposed for supervised ensembles of learning machines. Anyway, with clustering we have no univocally determined labels, and we need to refer to a "main" clustering (e.g. a clustering in the original high dimensional input space), to obtain a set of "reference" labels. Even this approach is in principle feasible, noise may be likely introduced, if the main clustering is too inaccurate. Here we adopt a combination scheme similar to that proposed by [Dudoit and Fridlyand, 2003], using the ensemble of clusterings to build a similarity matrix, and applying a second-level clustering algorithm to the lines of the matrix.

The *similarity matrix* M associated to a clustering $C =< A_1, A_2, \ldots, A_k >$ is a $n \times n$ matrix such that:

$$M_{ij} = \frac{1}{k} \sum_{s=1}^{k} I(i \in A_s) \cdot I(j \in A_s) \qquad (4)$$

where I is is the characteristic function of the set A_s, that is: if $i \in A_s$, $I(i \in A_s) = 1$, otherwise $I(i \in A_s) = 0$. The algorithm *RE-Clust* calls two clustering algorithms $\mathcal{C}$ and $\mathcal{C}_{com}$ to respectively generate the multiple clusterings and to combine the clustering results through the similarity matrix M. The high level pseudo-code of the ensemble algorithm scheme is the following:

***RE-Clust* algorithm**:

`Input`:

- a data set $X = \{x_1, x_2, \ldots, x_n\}$, represented by a D $d \times n$ matrix.
- an integer k (number of clusters)
- a real $\epsilon > 0$ (distortion level)
- an integer c (number of clusterings)
- two clustering algorithms $\mathcal{C}$ and $\mathcal{C}_{com}$
- a procedure that realizes a randomized map μ

`begin algorithm`

(1) Set the d' dimension of the projected subspace according to the *JL lemma*:
$d' = 2 \cdot \left(\frac{2 \log n + \log c}{\epsilon^2} \right)$

(2) Initialize the similarity matrix:
`For each` $i, j \in \{1, \ldots, n\}$ `do`
$M_{ij} = 0$

(3) `Repeat for` $t = 1$ `to` c

(4) Generate a realization P_t of the randomized map μ

(5) Generate the projected data D_t:
$D_t = P_t \cdot D$

(6) Apply the clustering algorithm $\mathcal{C}$ to D_t:
$< C_1^{(t)}, C_2^{(t)}, \ldots, C_k^{(t)} >= \mathcal{C}(D_t, k)$

(7) Generate the similarity matrices $M^{(t)}$:
`For each` $i, j \in \{1, \ldots, n\}$
$M_{ij}^{(t)} = \frac{1}{k} \sum_{s=1}^{k} I(i \in C_s^{(t)}) \cdot I(j \in C_s^{(t)})$

`end repeat`

(8) Compute M, the main similarity matrix:
$M = \frac{\sum_{t=1}^{c} M^{(t)}}{c}$

(9) Apply the clustering algorithm $\mathcal{C}_{com}$ to M to obtain the final clustering:

$$< A_1, A_2, \ldots, A_k >= \mathcal{C}_{com}(M, k)$$

`end algorithm`.

`Output`:

- the final clustering $C =< A_1, A_2, \ldots, A_k >$
- the similarity matrix M.

In the first step of the algorithm the dimension d' for the compressed data is computed. Since the failure probability is $e^{\Omega(-d'\epsilon^2)}$ (see Sect. 1.2), considering the c realizations $P_1, \ldots P_c$ of the randomized embedding μ (step 4) inside the repeat loop, we have, by union bound, that the following property holds:

$$P\left\{\forall x, y \in X, 1 \leq t \leq c, \left(\frac{1}{1+\epsilon} \leq \frac{||P_t x - P_t y||}{||x-y||} \leq 1+\epsilon\right)\right\} \geq 1 - cn^2 e^{\Omega(-d'\epsilon^2)}$$

Therefore, in the case of $\mu = PMO$, for $d' = 2\frac{2\log|X|+\log c}{\epsilon^2}$, with high probability we have that all the projections preserve the distances between the elements in X, up to a distortion $1+\epsilon$.

Inside the main repeat loop (step 3-7) a projected data set $D_t = P_t \cdot D$ is computed, the corresponding clustering $< C_1^{(t)}, C_2^{(t)}, \ldots, C_k^{(t)} >$ is obtained by calling $\mathcal{C}$, and a $M^{(t)}$ similarity matrix is built. After step (8), M_{ij} denotes the frequency by which the examples i and j occur in a randomly drawn cluster across multiple clusterings. If the k-clustering determines a partition, it is easy to see that $0 \leq M_{ij} \leq 1/k$. The final clustering is performed by applying the clustering algorithm $\mathcal{C}_{com}$ to the similarity matrix M.

We may choose different random projections to generate different *RE-Clust* ensembles. For instance, in our experiments (Sect. 3) we applied both *PMO* and *RS* random projections, to obtain the corresponding *PMO* and *RS* cluster ensembles.

3 Experiments with DNA microarray data

In this section we apply the proposed *RE-Clust* ensemble algorithm to gene expression data. The Ward's hierarchical agglomerative clustering algorithm [Ward, 1963] has been applied for both the base $\mathcal{C}$ and combining clustering algorithm $\mathcal{C}_{com}$, using as dissimilarity function the euclidean distance.

3.1 Experimental environment

We considered two DNA microarray data sets available on the web. The first one (*DLBCL-FL* data set) is composed by tumor specimens from 58 Diffuse Large B-Cell Lymphoma (DLBCL) and 19 Follicular Lymphoma (FL) patients [Shipp et al., 2002]. The second one (*Primary-Metastasis* data set) contains expression values in Affymetrix's scaled average difference units for 64 primary adenocarcinomas and 12 metastatic adenocarcinomas (lung, breast, prostate, colon, ovary, and uterus) from unmatched patients prior to any treatment [Ramaswamy et al., 2003]. In both cases we followed the same preprocessing and normalization steps described in [Shipp et al., 2002] and [Ramaswamy et al., 2003].

We compared classical single hierarchical clustering algorithm with our ensemble approach considering *PMO* and *RS* random projections (Sect. 1.3).

For each ensemble we randomly repeated the randomized projections 30 times, and each time we built *PMO* and *RS* ensembles composed by 50 base clusterings, for different ϵ values between 0.1 and 0.4.

Since clustering does not univocally associate a label to the examples, but only provides a set of clusters, we evaluated the error by choosing for each clustering the permutation of the classes that best matches the "a priori" known "true"

classes.

3.2 Results

Fig. 1 show the distributions of the errors of *RE-Clust* ensembles with DNA microarray data across 30 replications of the experiments.

In both cases *RE-Clust* ensembles perform equal or better than single hierarchical clustering, at least when relatively low distorted embeddings are chosen.

With the *DLBCL-FL* data set there is no significant difference in the accuracy achieved by *RE-Clust* ensembles and single hierarchical clustering (Fig. 1 a). In both cases we obtain an error of about 10 %. However note that, using Golub's weighted voting [Golub et al., 1999] and leave-one-out estimation of the error, [Shipp et al., 2002] achieved an error of about 9 %. Hence using an unsupervised method that does not use "a priori" knowledge on the data we obtain an error comparable with the one obtained by a supervised approach that on the contrary exploits the knowledge about the labels. This fact suggests that with *DLBCL-FL* probably is very difficult to lower the 10 % error using an unsupervised method, and hence in this case ensembling cannot improve the overall performance. Anyway, also when it is hard to improve the performance, we may apply ensembles to confirm the reliability of the discovered clusters. Indeed, we obtained the same error with all the 30 random repetitions of *PMO* ensembles each one composed by 50 base clusterings over 3499-dimensional random projections obtained from the original 6285-dimensional space ($\epsilon = 0.1$, Fig. 1 a).

Fig. 1 (b) shows that with *Primary-Metastasis* DNA microarray data [Ramaswamy et al., 2003], *Re-Clust* ensembles can improve single hierarchical clustering. Also with this data set we try to separate two classes (primary from metastatic tumors), but in this case the task is more difficult, because the primary tumors are heterogeneous, collecting lung, breast, prostate, colon, ovary, and uterus samples, as well as their metastatic counterparts.

In particular with *PMO* ensembles for different low distorted subspaces we obtain better results than that achieved with single hierarchical clustering (Fig. 1 b).

We obtained similar results also using the *Partitioning Around Medoids (PAM)* method [Kaufman and Rousseeuw, 1990] as base and combination clustering algorithm: with both data sets *PMO* and *RS* ensembles achieved equal or better results than single PAM clustering (data not shown).

It is worth noting that with both data sets *PMO* ensembles provide more stable results than *RS* ensembles, with a significantly lower dispersion of the error across the repeated experiments (Fig. 1).

4 Conclusions

We extended the *RS* approach to more general randomized projections that satisfy the *JL lemma* (Sect. 1.2), introducing a corresponding family of new ensemble clustering methods (*PMO, RS, Achlioptas*) based on randomized embeddings.a

Moreover we proposed a principled way to choose the distribution of the projected subspace, according to the *JL lemma*.

RE-Clust ensembles may improve the accuracy and the reliability of clustering algorithms, and are well-suited to the unsupervised analysis of DNA microarray data. Indeed *RE-Clust* ensembles can improve single clustering when a set of redundant features is involved, and this is exactly the case of gene expression data clustering problems.

An ongoing development of this work consists in a fuzzy extension of our algorithmic scheme. Indeed by substituting the characteristic function with a fuzzy or possibilistic membership, and the algebraic product with a more general *t-norm* (step 7, Sect. 2), we can obtain a fuzzy or a pos-

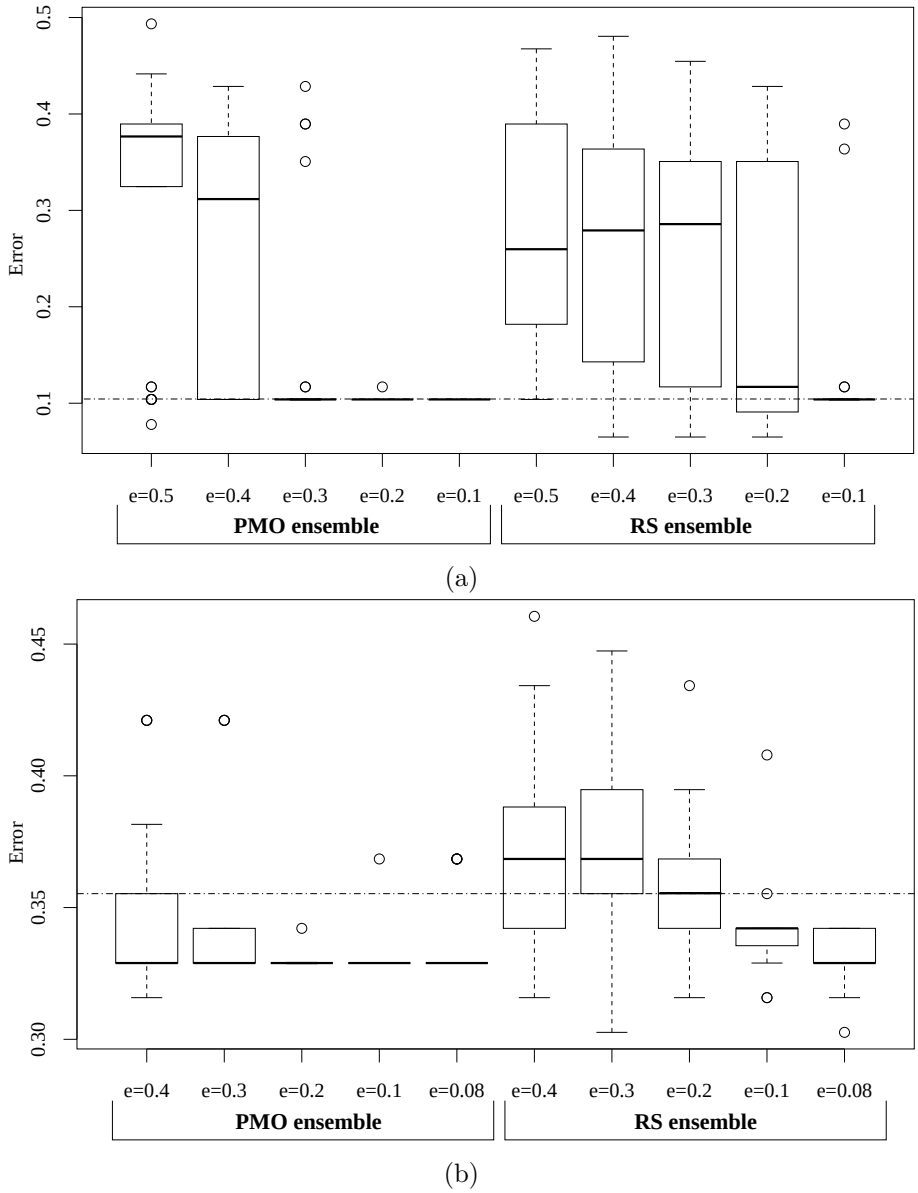

Figure 1: Boxplots of the errors with DNA microarray data. Single hierarchical clustering, PMO and RS ensembles with different $1+\epsilon$ distortions are compared, using 30 replications for each experiment, and 50 clusterings for each ensemble. The thin lines inside the boxes represent the median value. The error achieved with single hierarchical clustering is represented by the horizontal dash-dotted line. (a) *DLBCL-FL* data set (b) *Primary-Metastasis* DNA microarray data.

sibilistic version of the *Re-Clust* algorithm.

Acknowledgement

The present work has been developed in the context of the *CIMAINA* Center of Excellence, and it was partially funded by the italian COFIN project *Linguaggi formali ed automi: metodi, modelli ed applicazioni.*

References

D. Achlioptas. Database-friendly random projections. In P. Buneman, editor, *Proc. ACM Symp. on the Principles of Database Systems*, Contempo-

rary Mathematics, pages 274–281, New York, NY, USA, 2001. ACM Press.

R. Bellman. *Adaptive Control Processes: a Guided Tour*. Princeton University Press, New Jersey, 1961.

E. Bingham and H. Mannila. Random projection in dimensionality reduction: Applications to image and text data. In *Proc. of KDD 01*, San Francisco, CA, USA, 2001. ACM.

S. Dudoit and J. Fridlyand. A prediction-based resampling method for estimating the number of clusters in a dataset. *Genome Biology*, 3(7):1–21, 2002.

S. Dudoit and J. Fridlyand. Bagging to improve the accuracy of a clustering procedure. *Bioinformatics*, 19(9):1090–1099, 2003.

T.S. Furey, N. Cristianini, N. Duffy, D. Bednarski, M. Schummer, and D. Haussler. Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics*, 16(10):906–914, 2000.

T.R. Golub et al. Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. *Science*, 286:531–537, 1999.

I. Guyon and A. Elisseeff. An Introduction to Variable and Feature Selection. *Journal of Machine Learning Research*, 3:1157–1182, 2003.

T.K. Ho. The random subspace method for constructing decision forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(8):832–844, 1998.

X. Hu and I. Yoo. Cluster ensemble and its applications in gene expression analysis. In *Proc. 2nd Asia-Pacific Bioinformatics Conference*, pages 297–302, Dunedin, New-Zealand, 2004.

W.B. Johnson and J. Lindenstrauss. Extensions of Lipshitz mapping into Hilbert space. In *Conference in modern analysis and probability*, volume 26 of *Contemporary Mathematics*, pages 189–206. Amer. Math. Soc., 1984.

L. Kaufman and P.J. Rousseeuw. *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley, New York, 1990.

S. Monti, P. Tamayo, J. Mesirov, and T. Golub. Consensus Clustering: A Resampling-based Method for Class Discovery and Visualization of Gene Expression Microarray Data. *Machine Learning*, 52: 91–118, 2003.

S. Ramaswamy, K. Ross, E. Lander, and T. Golub. A molecular signature of metastasis in primary solid tumors. *Nature Genetics*, 33:49–54, 2003.

A. Rosenwald, R. Wright, W. Chan, J.M. Connors, R.I. Campo, E.and Fisher, R.D. Gascoyne, H.K. Muller-Hermelink, E.B. Smeland, J. Giltnane, E. Hurt, H. Zhao, L. Averett, L. Yang, W. Wilson, E. Jaffe, R. Simon, R.D. Klausner, J. Powell, P.L. Duffey, D.L. Longo, T.C. Greiner, D. Weisenburger, W.G. Sanger, B. Dave, J. Lynch, J. Vose, J. Armitage, E. Montserrat, A. Lopez-Guillermo, T. Grogan, T. Miller, M. LeBlanc, G. Ott, S. Kvaloy, J. Delabie, H. Holte, P. Krajci, T. Stokke, and L.M. Staudt. The use of molecular profiling to predict survival after chemoterapy for diffuse large-b-cell lymphoma. *New England Journal of Medicine*, 346(25):1937–1947, 2002.

M. Shipp, K. Ross, P. Tamayo, A. Weng, J. Kutok, R. Aguiar, M. Gaasenbeek, M. Angelo, M. Reich, G. Pinkus, T. Ray, M. Koval, K. Last, A. Norton, T. Lister, J. Mesirov, D. Neuberg, E. Lander, J. Aster, and T. Golub. Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. *Nature Medicine*, 8(1):68–74, 2002.

M. Smolkin and D. Gosh. Cluster stability scores for microarray data in cancer studies. *BMC Bioinformatics*, 36(4), 2003.

J.H. Ward. Hierarchical grouping to optimize an objective function. *J. Am. Stat. Assoc.*, 58:236–244, 1963.