Random projections for assessing gene expression cluster stability

Alberto Bertoni
DSI, Dipartimento di Scienze dell' Informazione,
Università degli Studi di Milano,
Via Comelico 39, Milano, Italy
E-mail: bertoni@dsi.unimi.it

Giorgio Valentini
DSI, Dipartimento di Scienze dell' Informazione,
Università degli Studi di Milano,
Via Comelico 39, Milano, Italy
E-mail: valentini@dsi.unimi.it

***Abstract*— Clustering analysis of gene expression is characterized by the very high dimensionality and low cardinality of the data, and two important related topics are the validation and the estimate of the number of the obtained clusters. In this paper we focus on the estimate of the stability of the clusters. Our approach to this problem is based on random projections obeying the Johnson-Lindenstrauss lemma, by which gene expression data may be projected into randomly selected low dimensional subspaces, approximately preserving pairwise distances between examples. We experiment with different types of random projections, comparing empirical and theoretical distortions induced by randomized embeddings between euclidean metric spaces, and we present cluster-stability measures that may be used to validate and to quantitatively assess the reliability of the clusters obtained by a large class of clustering algorithms. Experimental results with high dimensional synthetic and DNA microarray data show the effectiveness of the proposed approach.**

I. Introduction

Clustering methods may discover gene expression signatures related to specific biological processes or to specific diseases. Moreover unsupervised learning methods, exploiting the overall gene expression profile of a patient, may research and discover subclasses of pathologies that cannot be detected with traditional biochemical, histopathological and clinical criteria [4].

Two of the main concerns with gene expression clustering analysis are the estimate of the number of clusters in a dataset, and the stability of the obtained clusters [3]. Indeed in many cases we have no sufficient biological knowledge to "a priori" evaluate both the number of clusters (e.g. the number of biologically distinct tumor classes), as well as the validity of the discovered clusters (e.g. the reliability of new discovered tumor classes).

Several approaches for assessing the reproducibility and stability of clustering patterns in gene expression data have been recently proposed [8], [9], [13].

In this paper we present an approach that exploits the very high dimensionality and relatively low cardinality of gene expression data, using multiple random projections of the original data, to assess the reliability of the discovered clusters. The main idea behind our approach consists in evaluating the stability of the clusters discovered in the original high dimensional space comparing them with the clusters discovered in randomly projected lower dimensional subspaces. To this end we use the concept of random projections with bounded metric distortions, according to the Johnson-Lindenstrauss (*JL*) theory [7].

The proposed method is related to the Smolkin and Gosh [12] approach based on an unsupervised version of the random subspace method [5]. We extend the unsupervised random subspace approach to more general random projections, in the framework of random embeddings between euclidean spaces, and we propose a new cluster stability measure based on similarity between randomly projected data.

In the next section we present a brief introduction to randomized embeddings of metric spaces, focusing on random projections obeying the *JL* lemma. In Sect. III we compare the theoretical and empirical distortion induced by randomized embeddings using two high-dimensional synthetic data. Then in Sect. IV we present our approach to the estimate of cluster stability based on random projections, and we apply the proposed stability measures to both synthetic and "real" gene expression data.

For all the experiments presented in this paper we developed *R* functions and programs to implement both the random projections described in Sect.III and the stability measures described in Sect. IV.

II. Dimensionality reduction and randomized embeddings

Dimensionality reduction may be obtained by mapping points from a high to a low-dimensional space, approximately preserving some characteristics, i.e. the distances between points. In this context randomized embeddings with low distortion represent a key concept. Randomized embeddings have been successfully applied both to combinatorial optimization and data compression [6].

A *randomized embedding* between L_2 normed metric spaces with distortion $1+\epsilon$, with $\epsilon > 0$ and failure probability P is a distribution probability over mappings $\mu : \mathbb{R}^d \rightarrow \mathbb{R}^{d'}$, such that for every pair $p, q \in \mathbb{R}^d$, the following property holds with probability $1 - P$:

$$\frac{1}{1+\epsilon} \leq \frac{||\mu(p) - \mu(q)||_2}{||p - q||_2} \leq 1 + \epsilon \qquad (1)$$

The main result on randomized embedding is due to Johnson and Lindenstrauss [7], who proved the existence of a randomized embedding $\mu : \mathbb{R}^d \rightarrow \mathbb{R}^{d'}$ 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:

$$Prob\left(\frac{1}{1+\epsilon} \leq \frac{||\mu(p)-\mu(q)||_2}{||p-q||_2} \leq 1+\epsilon\right) \geq 1-n^2 e^{\Omega(-d'\epsilon^2)} \quad (2)$$

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 [7] 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, 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. In particular in [1] matrices are proposed such that their entries are uniformly chosen in $\{-1, 1\}$, or in $\{-\sqrt{3}, 0, \sqrt{3}\}$, by choosing 0 with probability $2/3$ and $-\sqrt{3}$ or $\sqrt{3}$ with probability $1/6$. In this case the *JL lemma* holds with $c \simeq 4$.

Consider now a data set represented by a $d \times n$ matrix X whose columns represent n d-dimensional observations. Suppose that $d' = 4\ \log n/\epsilon^2 << d$; the *JL lemma* guarantees the existence of a $d' \times d$ matrix P such that the columns of the "compressed" data set $X^P = PX$ have approximately the same distance (up to a distortion $1+\epsilon$) of the corresponding columns in X. Moreover there is a randomized algorithm that, having in input X, outputs X^P in time $\mathcal{O}(dd'n)$ with high confidence.

This fact suggests that we can speed-up algorithms for solving *proximity problems*. Instances of a *proximity problem* are sets $I \subset \mathbb{R}^d$ (described by a data set X), and the goal consists in computing some properties defined in terms of distances between points in I: clustering is an example. In particular consider an algorithm $\mathcal{A}$ that, having as input a $d \times n$ data set X, outputs the solution of a *proximity problem* in time $T(n, d)$. An approximate solution of the problem can be obtained by computing firstly the projection P and the "compressed" data set $X^P = PX$, and finally by applying $\mathcal{A}$ to X^P. In this way the time complexity may be reduced from $T(n, d)$ to $\mathcal{O}(nd \log n) + T(n, \mathcal{O}(\log n))$.

III. Distortion induced by random projections

In this section we consider two random embeddings, proposed respectively in [1] and [5]. We estimate the distortions induced by the random embeddings with respect to high dimensional synthetic data, comparing them with the theoretical bounds predicted by the *JL lemma*.

A. Distortion measures

Given a data set $X \subset \mathbb{R}^d$ and a map $\mu : \mathbb{R}^d \rightarrow \mathbb{R}^{d'}$, for $x, y \in X$ the *distortion* $dist_\mu(x, y)$ is defined:

$$dist_\mu(x, y) = \frac{||\mu(x)-\mu(y)||_2}{||x-y||_2} \quad (3)$$

Of course, $dist_\mu(x, y) = 1$ means that no distortion is introduced. The *maximum*, *minimum* and *average distortion* of μ on X respectively are:

$$\begin{aligned} max.dist_\mu(X) &= \max_{x,y \in X} dist_\mu(x, y) \quad (4) \\ min.dist_\mu(X) &= \min_{x,y \in X} dist_\mu(x, y) \\ ave.dist_\mu(X) &= \frac{1}{|X|(|X|-1)} \sum_{x,y \in X, x \neq y} dist_\mu(x, y) \end{aligned}$$

B. Empirical estimation of distortions induced by randomized maps

In this section we estimate, given a data set X and a randomized map μ, the expectation of the random variables $max.dist_\mu(X), min.dist_\mu(X)$ and $ave.dist_\mu(X)$ (eq.4)

1) Randomized maps: We considered two randomized maps:

- *Random Projection (RP)*: represented by $d' \times d$ matrices $P = 1/\sqrt{d'}(r_{ij})$, where r_{ij} are uniformly chosen in $\{-1, 1\}$. As observed in Sect.II *RP* satisfies the *JL lemma*.
- *Random Subspace (RS)* [5]: 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 for a $d \times n$ data set X and a projection matrix P, the "compressed" data set $X^P = PX$ can be computed in time $\mathcal{O}(nd')$, independently from d. Unfortunately, *RS* does not satisfy the *JL lemma*.

2) Synthetic data generation: We developed two generators for synthetic data sets (*sample1* and *sample2*):

- *Sample1* is a generator for 6000-dimensional data sets composed by 3 clusters of data normally distributed. The elements of each cluster are distributed according to a spherical gaussian with unitary standard deviation. The first cluster is centered in the middle of a 6000-dimensional hypercube with an edge of length equal to 20 conventional units. The other two clusters are centered at the opposite vertices of the hypercube. Hence the tree clusters are completely separated with no overlapping between them.
- *Sample2* is a a generator for 6000-dimensional data sets composed by 5 clusters of data normally distributed. All the examples have 1000 no-noisy and 5000 noisy variables; for all the examples the noisy variables are distributed according to a spherical gaussian centered in 0 and with standard deviation equal to 2. Considering only the 1000 no-noisy variables there is substantial overlapping between classes 1 and 2 and 1 and 3, while class 4 and 5 are quite well separated.

Using the generators we drew two data set (respectively X_1 and X_2), each one composed by 50 examples.

3) Results: Setting a distortion value $1+\epsilon$, $(0 < \epsilon < 0.5)$, a dimension $d' = 4\ \log 50/\epsilon^2$ is computed according to the *JL lemma*. For every data set X_1 and X_2 we performed 50 *RP* and 50 *RS* projections, computing the empirical average of $max.dist_\mu, min.dist_\mu$ and $ave.dist_\mu$, according to eq.4.

The results for *RP* and *RS* on *sample2* are summarized in Fig.1. As expected, for *RP* the empirical average of $max.dist_\mu$ and $min.dist_\mu$ are significantly better than the theoretical bound. Quite surprisingly, similar results have been also obtained with *RS* projections, where *JL* bounds are not guaranteed. A similar behaviour of *RP* and *RS* projections has been also observed with *sample1* (data not shown).

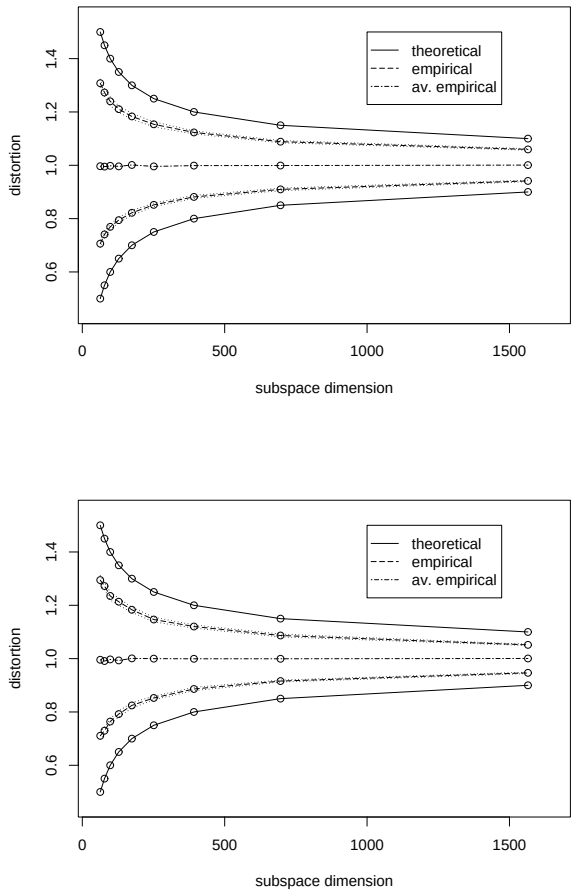

Fig. 1. Comparing theoretical and empirical distortion with *sample2* using *RP* and *RS* projections. Continuous lines represent the bounds of the maximum and minimum distortion according to the *JL* lemma. Dashed lines represent the average maximum and minimum distortion empirically computed and averaged over 50 random projections. The pairs of dotted lines just above and below the the dashed lines represent the confidence interval at 99 % confidence level. The dash-dotted line represents the expected average distortion. Above: *RP* projection. Below: *RS* projection.

Fig. 2 shows the distribution of the pairwise distances between examples in original and randomly projected data (*sample2* data set). We may see that the also the distributions of the pairwise distances are quite well preserved, at least if we project data with low distortion. With the well separated clusters of the *sample1* data set the distribution of the distances are better preserved (data not shown).

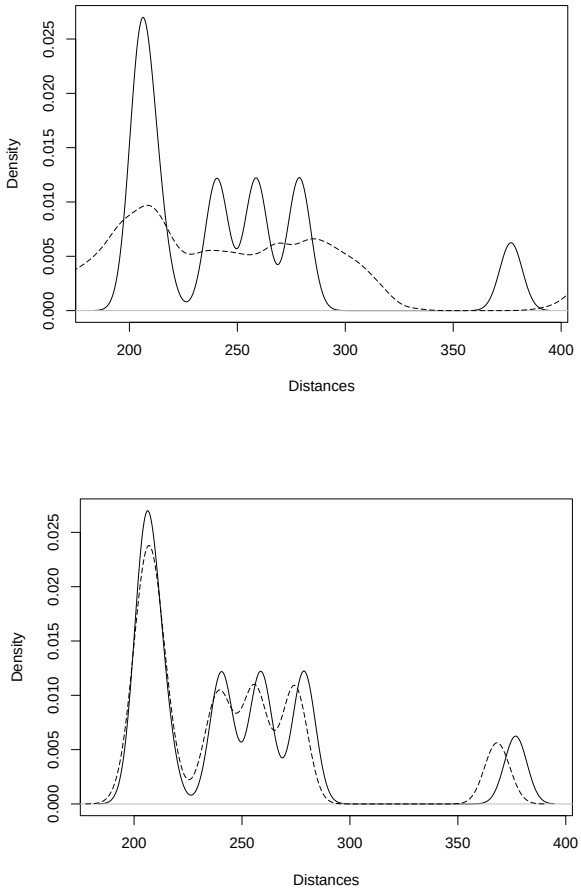

Fig. 2. Distribution of the pairwise distances between examples in original and randomly projected data (*sample2*). The continuous line represents the distribution in the original 6000-dimensional space, the dashed line the distribution in the projected space. Above: Projection into a 63-dimensional space (corresponding to a 1.50 upper-bound distortion according to JL lemma). Below: Projection into a 1565-dimensional space (corresponding to a 1.10 upper-bound distortion according to JL lemma).

IV. Random projections and cluster stability

The *JL* lemma shows that we may generate relatively low-distorted random projected data, and our experimental results show that we may also obtain empirical estimate of the expectation of the random variables $max.dist_\mu$ and $min.dist_\mu$ that are better than the theoretical bounds.

Our aim is to exploit random projections to estimate stability of clusters, because random projections do not induce relevant distortions (as long as we provide a projection into a sufficiently high-dimensional subspace).

A. Cluster stability measures

Given a finite set $X \subset \mathbb{R}^d$, we denote (with abuse of notation) with X the metric space $< X, f >$, where $f(x,y) = ||x-y||_2,\ x,y \in \mathbb{R}^d$. In the following of this section we consider a fixed random projection $\mu : \mathbb{R}^d \to \mathbb{R}^{d'}$ that verifies the *JL* lemma (i.e. *RP*, Sect. III-B.1), and we propose a stability index for clustering by using a pairwise similarity matrix between the projected examples.

Let $\mathcal{C}$ be a clustering algorithm, that, having in input X, outputs a set of k clusters:

$$\mathcal{C}(X) =< A_1, A_2, \ldots, A_k >,\ A_j \subset X, 1 \leq j \leq k \quad (5)$$

Then we compute a "similarity" matrix M, with indices in X, using the following algorithm:

1) Generate t independent projections $\mu_i : \mathbb{R}^d \to \mathbb{R}^{d'},\ 1 \leq i \leq t$, such that $d' = 4\,\frac{\log|X| + \log t}{\epsilon^2}$
2) Apply $\mathcal{C}$ to the new projected data $\mu_i(X)$, obtaining a set of clusterings, for $1 \leq i \leq t$:
$$\mathcal{C}(\mu_i(X)) =< B_1^i, \ldots, B_k^i >,\ B_j^i \subset X_i, 1 \leq j \leq k \quad (6)$$
where B_j^i is the j^{th} cluster of the i^{th} clustering.
3) Set the elements M_{xy} of the similarity matrix:
$$M_{xy} = \frac{1}{t}\sum_{j=1}^{k}\sum_{i=1}^{t} \chi_{B_j^i}(\mu_i(x)) \cdot \chi_{B_j^i}(\mu_i(y)) \quad (7)$$
where $\chi_{B_j^i}$ is the characteristic function for the cluster B_j^i.

Since the elements M_{xy} measure the occurrences of the examples $\mu_i(x), \mu_i(y) \in \mu_i(X)$ in the same clusters B_j^i for $1 \leq i \leq t$, then M represents the "tendency" of the projections to belong to the same cluster. It is easy to see that $0 \leq M_{xy} \leq 1$, for each $x, y \in X$.

With respect to the algorithm above we may observe:

Remark 1. Since the failure probability is $e^{\Omega(-d'\epsilon^2)}$, similarly to eq.2 in Sect. II, by union bound we have, for all $x, y \in X$, $1 \leq i \leq t$:

$$P\left(\frac{1}{1+\epsilon} \leq \frac{||\mu_i(y) - \mu_i(x)||_2}{||x-y||_2} \leq 1+\epsilon\right) \geq 1 - t|X|^2 e^{\Omega(-d'\epsilon^2)}$$

Therefore for $d' \simeq \mathcal{O}\left(\frac{\log|X| + \log t}{\epsilon^2}\right)$, we obtain with high probability that all the projections preserve the distances between the elements in X up to a distortion $1+\epsilon$.

Remark 2. A fuzzy similarity matrix may be obtained simply substituting in eq. 7 the characteristic function with a membership function and the algebraic product with a suitable *t-norm*. In this way fuzzy or possibilistic clustering approaches may also be applied.

Using the similarity matrix M (eq. 7) we propose the following *stability index* s for a cluster A_i:

$$s(A_i) = \frac{1}{|A_i|(|A_i|-1)} \sum_{(x,y)\in A_i \times A_i, x \neq y} M_{xy} \quad (8)$$

The index $s(A_i)$ estimates the stability of a cluster A_i in the original non projected space, by measuring how much the projections of the pairs $(x,y) \in A_i \times A_i$ occur together in the same cluster in the projected subspaces. The stability index has values between 0 and 1: values near 1 denote stable clusters, while lower values indicate less reliable clusters. The above stability index is very similar to that proposed by [10]. The main difference of our approach consists in the way the similarity matrix is computed: we applied randomized projections into lower dimensional subspaces, while [10] applied bootstrap techniques.

An overall measure of the stability of the clustering in the original space may be obtained averaging between the stability indices:

$$S(k) = \frac{1}{k}\sum_{i=1}^{k} s(A_i) \quad (9)$$

In this case also we have that $0 \leq S(k) \leq 1$, where k is the number of clusters.

B. Assessing cluster stability in synthetic and gene expression data

We applied the stability measures proposed in the previous section to high dimensional synthetic and gene expression data, using the Ward's hierarchical agglomerative clustering algorithm [14], and using as dissimilarity function the euclidean distance.

For each data set we computed the average stability index $S(k)$ (eq. 9) for different number k of clusters, and the stability index s (eq. 8) for each corresponding cluster, considering different $1+\epsilon$ distortions induced by *RS* and *RP* projections (Sect. III-B.1) into subspaces whose dimension was computed according to the *JL lemma*.

1) Results with synthetic data: Tab.I summarizes the results with *sample1*. The maximum of the average stability index $S(k)$ is reached when the dendrogram is cut at 3 clusters level, and the corresponding stability indices s are equal to 1 for each of the 3 clusters. Both the average and the individual stability indices are lower when different number of clusters are selected, showing that the proposed stability measures correctly detect 3 clusters, identifying them as highly reliable.

With *sample2* the stability indices correctly predict largely separated as well as less reliable clusters. Indeed the stability indices are high for the 2 well separated clusters, while for the other overlapped clusters the stability indices are significantly lower (data not shown).

2) Results with gene expression data data: We applied the proposed stability indices to a set of gene expression tumor specimens from 58 Diffuse large B-cell lymphoma (DLBCL) and 19 Follicular lymphoma (FL) patients [11].

Tab. II shows the estimate of cluster stability for the *DLBCL-FL* data set. Note that in the first column of Tab. II the clusters are labeled with numbers, and these number assignments correspond to left-to-right clusters in the dendrogram of Fig. 3. The average S index is slightly larger when the hierarchical clustering dendrogram is cut at 2 clusters level (Fig. 3), but comparable (even if lower) values are also registered with $3, 4$ and 5 clusters. In this case indeed the

TABLE I

Sample1: ESTIMATE OF CLUSTER STABILITY.

Clusters	**Members of Clusters**	**Stability index s** $\epsilon = 0.5$	$\epsilon = 0.4$	$\epsilon = 0.3$	$\epsilon = 0.2$	$\epsilon = 0.1$
2 clusters		$S = 0.8631$	$S = 0.8684$	$S = 0.8684$	$S = 0.9157$	$S = 0.9421$
1	11-20	1.0000	1.0000	1.0000	1.0000	1.0000
2	1-10,21-30	0.7263	0.7368	0.7368	0.8314	0.8842
3 clusters		$S = 1.0000$	$S = 1.0000$	$S = 1.0000$	$S = 1.0000$	$S = 1.0000$
1	11-20	1.0000	1.0000	1.0000	1.0000	1.0000
2	21-30	1.0000	1.0000	1.0000	1.0000	1.0000
3	1-10	1.0000	1.0000	1.0000	1.0000	1.0000
5 clusters		$S = 0.7059$	$S = 0.6843$	$S = 0.7044$	$S = 0.7004$	$S = 0.7472$
1	11,13,16,17,19,20	0.6973	0.7346	0.7293	0.6506	0.7560
2	12,14,15,18	0.6666	0.7066	0.6866	0.6466	0.7133
3	21-30	0.7155	0.7582	0.7448	0.7591	0.8364
4	5,7	0.7600	0.5600	0.6800	0.7400	0.7800
5	1-4,6,8-10	0.6900	0.6621	0.6814	0.7057	0.6507
10 clusters		$S = 0.3093$	$S = 0.3043$	$S = 0.2651$	$S = 0.3286$	$S = 0.3936$
1	19	0.0600	0.1200	0.0600	0.2000	0.2400
2	11,13,16,17,20	0.4260	0.3520	0.2900	0.3360	0.4560
3	12	0.1400	0.1600	0.1600	0.2000	0.1400
4	14,15,18	0.4066	0.3533	0.3200	0.3800	0.4200
5	23,28,29	0.3733	0.3000	0.2866	0.3600	0.4200
6	21,22,24-27,30	0.3276	0.3419	0.3285	0.3866	0.3933
7	5,7	0.3600	0.2800	0.3000	0.3600	0.3800
8	2,3,8,10	0.3000	0.3366	0.3066	0.3433	0.3866
9	4,9	0.3400	0.4000	0.2600	0.4200	0.5000
10	1,6	0.3600	0.4000	0.3400	0.3000	0.6000

clusters are not clearly delineated. For instance, considering a cut at 4 clusters level, the first cluster (with a relatively high s stability index equal to 0.8748) is composed by homogeneous FL patients (Fig. 3), the second (less reliable $s = 0.6004$) is composed by both DLBCL and FL patients, while the third (more reliable $s = 0.8123$) is composed only by DLBCL patients, as well as the less reliable ($s = 0.6005$) fourth cluster. Splitting the fourth cluster, we obtain two DLBCL subclusters, more reliable than the previous one (Tab. II, 5 clusters). If we split the data in 10 or more clusters we note a significant decrement of both the s indices and the average S index: this fact suggests that no significant structure can be observed in small-sized clusters (data not shown).

These results are congruent with the bio-medical characteristics of the data. Indeed even if nodal tumor specimens are subdivided into 2 groups (DLBCL and FL), Alizadeh et al. [2] discovered subclasses among DLBCL patients, and Shipp et al. [11] highlighted that FL patients frequently evolve over time and acquire the morphologic and clinical features of DLBCLs.

V. CONCLUSIONS

Our experiments with synthetic and gene expression data show that the proposed stability indices based on random projections with bounded metric distortion may be used to identify stable clusters directly from the data, without "a priori" knowledge and without assumptions about the distribution of the data (apart of the choice of the clustering algorithm). Moreover our experiments show that the average stability index may also be useful to identify the most likely number of clusters.

We experimented with agglomerative hierarchical clustering, but the proposed approach may be used with any clustering algorithm, comprising also fuzzy and possibilistic clustering methods.

Our experimental results show also that, according to the *JL* lemma, if the dimension of the subspace induced by a random projection is sufficiently high, no significant distortion is introduced into the embedding, and clustering may be performed on random subspaces approximately preserving pairwise distances between examples. From this standpoint, our random projection-based stability measures may help biomedical researchers to identify stable and reliable clusters (e.g. new pathological classes), exploiting the high dimension of gene expression data.

ACKNOWLEDGMENT

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

REFERENCES

[1] D. Achlioptas. Database-friendly random projections. In *Proc. ACM Symp. on the Principles of Database Systems*, Contemporary Mathematics, pages 274–281, 2001.

[2] Alizadeh, A. et al. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*, 403:503–511, 2000.

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

[4] M.B. Eisen, P.T. Spellman, P.O. Brown, and D. Botstein. Cluster analysis and display of genome-wide expression patterns. *PNAS*, 95(25):14863–14868, 1998.

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

[6] P. Indyk. Algorithmic Applications of Low-Distortion Geometric Embeddings. *Proceedings of the 42nd IEEE symposium on Foundations of Computer Science*, pp. 10-33, IEEE Computer Society, Washington DC, USA, 2001.

TABLE II

DLBCL-FL: ESTIMATE OF CLUSTER STABILITY.

Clusters	**Stability index s** $\epsilon = 0.5$	$\epsilon = 0.4$	$\epsilon = 0.3$	$\epsilon = 0.2$	$\epsilon = 0.1$
2 clusters	$S = 0.6620$	$S = 0.6433$	$S = 0.6624$	$S = 0.7140$	$S = 0.7826$
1	0.6998	0.6864	0.6893	0.7620	0.8936
2	0.6242	0.6002	0.6355	0.6660	0.6716
3 clusters	$S = 0.5369$	$S = 0.5303$	$S = 0.5720$	$S = 0.6655$	$S = 0.7536$
1	0.5258	0.5038	0.5222	0.5115	0.6474
2	0.6081	0.6197	0.6749	0.8419	0.9149
3	0.4767	0.4675	0.5190	0.6432	0.6986
4 clusters	$S = 0.4822$	$S = 0.4829$	$S = 0.5167$	$S = 0.6025$	$S = 0.7220$
1	0.5443	0.5392	0.6351	0.6828	0.8748
2	0.4949	0.4760	0.4496	0.4909	0.6004
3	0.5265	0.5327	0.5725	0.7340	0.8123
4	0.3633	0.3839	0.4094	0.5024	0.6005
5 clusters	$S = 0.4164$	$S = 0.4378$	$S = 0.4608$	$S = 0.5660$	$S = 0.6946$
1	0.4825	0.4979	0.5646	0.6335	0.8492
2	0.4281	0.4329	0.3995	0.4362	0.5257
3	0.4437	0.4621	0.5087	0.6417	0.7241
4	0.3443	0.3921	0.4165	0.5418	0.6275
5	0.3836	0.4038	0.4146	0.5769	0.7467

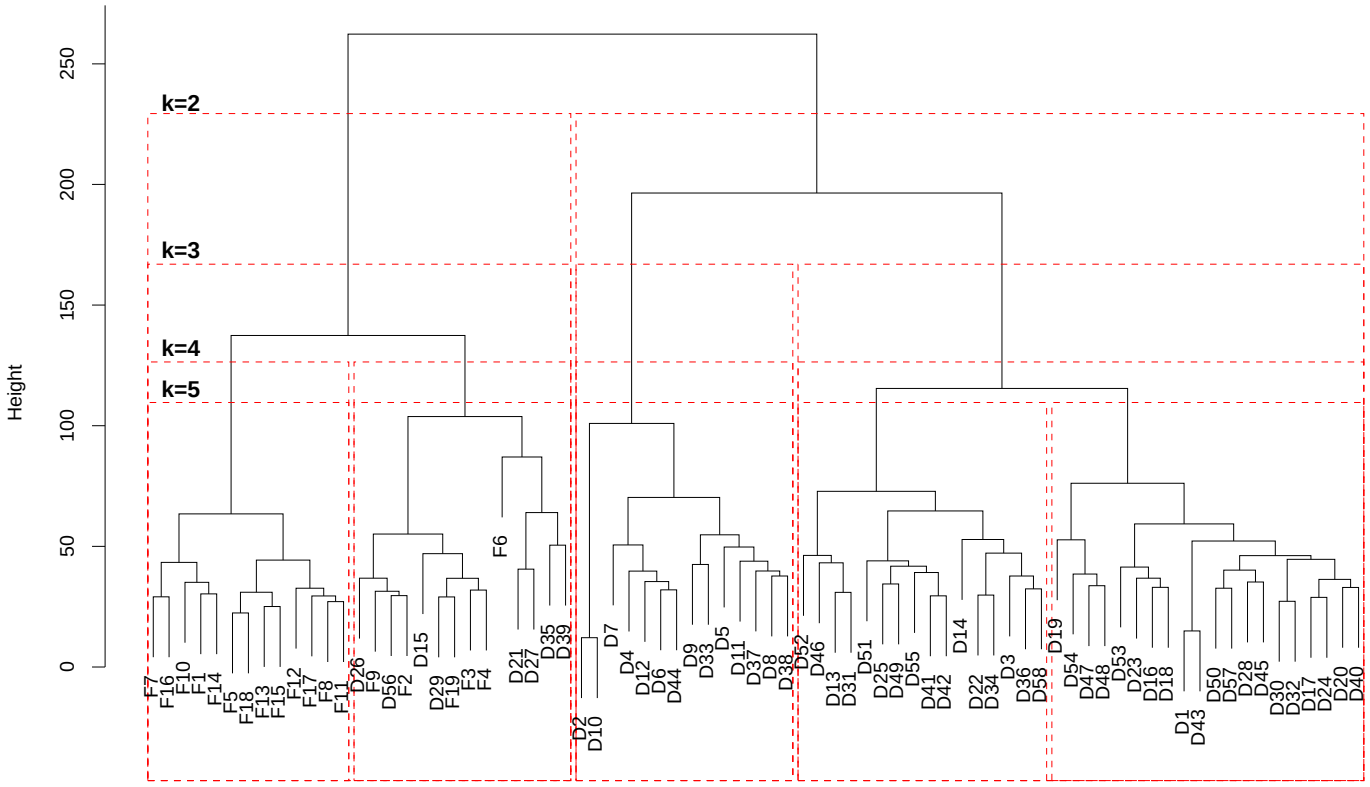

Fig. 3. Hierarchical clustering of *DLBCL-FL* examples (Ward method). Leaves labeled with "D" refer to DLBCL patients, while "F" to FL patients. Gray dotted lines cut the dendrogram such that exactly k clusters are produced, for $k = 2, 3, 4, 5$.

[7] 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.

[8] M.K. Kerr and G.A. Curchill. Bootstrapping cluster analysis: assessing the reliability of conclusions from microarray experiments. *PNAS*, 98:8961–8965, 2001.

[9] L.M. McShane, D. Radmacher, B. Freidlin, R. Yu, M.C. Li, and R. Simon. Method for assessing reproducibility of clustering patterns observed in analyses of microarray data. *Bioinformatics*, 18(11):1462–1469, 2002.

[10] S. Monti et al. Consensus Clustering: A Resampling-based Method for Class Discovery and Visualization of Gene Expression Microarray Data. *Machine Learning*, 52:91–118, 2003.

[11] M. Shipp et al. Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. *Nature Medicine*, 8(1):68–74, 2002.

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

[13] R. Tibshirani, T. Hastie, B. Narasimham, M. Eisen, G. Sherlock, P. Brown, and D. Botstein. Exploratory screening of genes and clusters from mictroarray experiments. *Statist. Sinica*, 12:47–60, 2002.

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