BIOINFORMATICS

Vol. 00 no. 00
Pages 1–2

Clusterv*: a tool for assessing the reliability of clusters discovered in DNA microarray data*

Giorgio Valentini

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

ABSTRACT

Summary: We present a new R package for the assessment of the reliability of clusters discovered in high dimensional DNA microarray data. The package implements methods based on random projections that approximately preserve distances between examples in the projected subspaces.

Availability: http://homes.dsi.unimi.it/∼valenti/SW/clusterv/download/clusterv_1.0.tar.gz

Contact: valentini@dsi.unimi.it

Supplementary Information:
http://homes.dsi.unimi.it/∼valenti/SW/clusterv

An open problem in microarray data analysis is the assessment of the reliability of clustering results. Indeed a quantitative data-driven estimate of the reliability of the discovered clusters can support bio-medical researchers in the validation of novel subgroups identified at bio-molecular level. Most of the proposed works focused on the estimate of the number of clusters (see e.g. Azuaje (2002)), but only few works proposed methods to assess the reliability of the individual clusters discovered using DNA microarray data (McShane et al., 2002; Smolkin and Gosh, 2003).

Our approach exploits the redundancy of features (gene expression levels) that characterize DNA microarray data. Exploiting this redundancy, we extend to more general low-distorted random projections a previous approach based on random subspace methods (Smolkin and Gosh, 2003). We quantitatively evaluate the reliability of the discovered clusters by using multiple random projections of the original high dimensional data to lower dimensional subspaces, approximately preserving the distances between examples, according to the *Johnson-Lindenstrauss theory* (Johnson and Lindenstrauss, 1984). Our concept of reliability is tied to the concept of stability: comparing the clusters obtained by using multiple instances of the randomly projected data with the clusters obtained in the original high dimensional gene space, we measure if and at which extent the individual clusters are stable, that is ”close” to that obtained in the projected subspaces (see Supplementary Information for more details).

To compute the stability measures implemented in our R package *clusterv* (that stands for *cluster*-*v*alidity), we used a $n \times n$ symmetric similarity matrix M, whose elements M_{ij} store the memberships of examples pairs i, j to the same cluster (Dudoit and Fridlyand, 2003):

$$M_{ij} = \sum_{s=1}^{k} \chi_{C_s}[i] \cdot \chi_{C_s}[j] \tag{1}$$

where $i, j \in \{1, 2, \ldots, n\}$, $C_s \subseteq \{1, 2, \ldots, n\}$ is a cluster returned by a clustering algorithm, k the number of clusters, and $\chi_{C_s} \in \{0, 1\}^n$ is the characteristic vector of C_s, i.e. $\chi_{C_s}[i] = 1$ if $i \in C_s$, otherwise $\chi_{C_s}[i] = 0$. In other words M_{ij} denotes if elements i and j belong to the same cluster. Using multiple random projections of the data we generate multiple instances of projected data that are used by a clustering algorithm to provide multiple sets of clusters (clusterings). We then build multiple similarity matrices (one for each clustering), and averaging between them, we obtain a similarity matrix $\overline{M}$ that stores the memberships of examples pairs i, j to the same cluster across multiple clusterings.

Using the previously computed similarity matrix, the *stability index* s for an individual cluster C is:

$$s(C) = \frac{1}{|C|(|C|-1)} \sum_{(i,j) \in C \times C, i \neq j} \overline{M}_{ij} \tag{2}$$

The index $s(C)$ estimates the stability of a cluster C by measuring how much the projections of the pairs $(i, j) \in C \times C$ occur together in the same cluster in the projected subspaces. The stability index has values between 0 and 1: low values indicate no reliable clusters, high values denote stable clusters. An overall measure of the stability of the clustering may be obtained averaging between the stability indices:

$$S(k) = \frac{1}{k} \sum_{r=1}^{k} s(C_r) \tag{3}$$

In this case also we have that $0 \leq S(k) \leq 1$, where k is the number of clusters. Finally, the *Assignment-Confidence (AC)* index estimates the confidence of the assignment of an example i to a cluster C, by measuring the frequency by which i

appears with the other elements of the cluster A:

$$AC(i, C) = \frac{1}{|C| - 1} \sum_{j \in C, j \neq i} \overline{M}_{ij} \qquad (4)$$

The *clusterv* R package implements the above stability measures and provides a set of functionalities to assess the reliability of clusters discovered in data characterized by high-dimensionality (Tab. 1). A reliability analysis of clusters discovered by hierarchical, k-means, fuzzy k-means or PAM clustering algorithms can be performed by calling a single high-level function. For instance the function `Random.fuzzy.kmeans.validity` applies the fuzzy-k-means clustering algorithm to the data, computes the similarity matrix using multiple random subspace projections and then computes the stability indices for each cluster, the overall stability index of the clustering and the set of AC indices for each example. Moreover, using the functions clustering-algorithm-independent, the same reliability analysis can be performed for any distance-based clustering algorithm, as long as its output (the clustering) can be coded as an R vector or a list. For instance, the function `Cluster.validity` accepts as input the list of the clusters whose validity indices need to be computed and a list of the sets of clusters obtained from the randomly projected subspaces, and returns the stability indices described by eq. 2, 3 and 4.

Table 1. *Clusterv*: main functionalities

Functions clustering-algorithm independent
1. Functions for high dimensional synthetic data generation
2. Functions to implement different types of random projections from high to lower dimensional subspaces
3. Functions to compute the similarity matrix (eq. 1)
4. Functions to compute the stability indices (eq. 2, 3, 4)
Functions clustering-algorithm dependent
1. Functions to perform multiple clusterings on multiple instances of projected data
2. Functions to compute the stability indices for a specific clustering algorithm

As an example of an application of the *clusterv* R package, we briefly present a reliability analysis of the clusters obtained in melanoma patients (Fig. 1), using a cDNA microarray data set of 31 examples (Bittner et al., 2000). The overall stability index (eq. 3) estimates as $N = 4$ the optimal number of clusters (Tab. 2). With 4 clusters the first two clusters are singletons, while the third is a big cluster with 23 examples, including the 19-members melanoma subclass found out by Bittner et al.; the fourth very stable cluster groups together the remaining 6 examples. To find the same 19-members Bittner's cluster we need to choose a partition with $N = 9$ clusters: the fifth cluster exactly corresponds to it. However the stability index of this cluster is quite low ($s \simeq 0.4$), suggesting that the reliability of the cluster would be higher if we add to it 4 more specimens. More examples of applications of the proposed stability measures, a theoretical background on low distorted random projections, as well as a tutorial and a reference manual for using the *clusterv* R package, are available at the *clusterv* web site (see Supplementary Information).

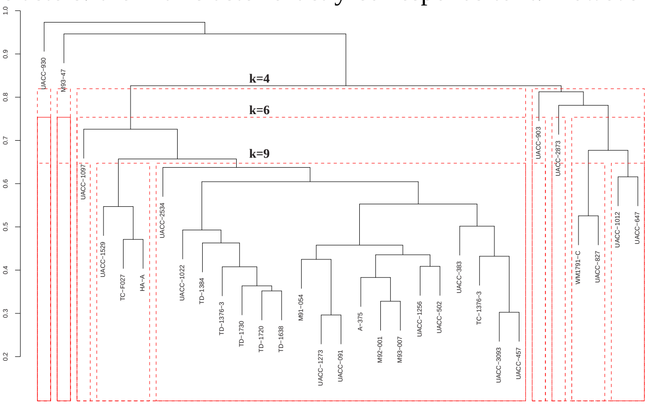

Fig. 1. Hierarchical clustering of *Melanoma* samples. Gray dotted lines cut the dendrogram such that exactly k clusters are produced, for $k = 4, 6, 9$. See Table 2 for the the corresponding stability indices.

Table 2. Cluster reliability analysis in *melanoma* patients.

Number of clusters:	4	5	6	7	8	9	10	12
Overall stability index S	0.97	0.73	0.59	0.50	0.55	0.52	0.54	0.39
Stability index s of individual clusters								
4 clusters:	1	2	3	4				
s	1.00	1.00	0.91	1.00				
5 clusters:	1	2	3	4	5			
s	1.00	1.00	0.66	0.00	1.00			
6 clusters:	1	2	3	4	5	6		
s	1.00	1.00	0.56	0.00	0.00	1.00		
9 clusters:	1	2	3	4	5	6	7	8	9
s	1.00	1.00	0.00	0.35	0.40	0.00	0.00	0.94	0.98

REFERENCES

F. Azuaje. A cluster validity framework for genome expression data. *Bioinformatics*, 18(2):319–320, 2002.

M. Bittner et al. Molecular classification of malignant melanoma by gene expression profiling. *Nature*, 406:536–540, 2000.

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

W.B. Johnson and J. Lindenstrauss. Extensions of Lipshitz mapping into Hilbert space. *Contemporary Mathematics*, 26, pages 189–206. Amer. Math. Soc., 1984.

L.M. McShane, et al. Method for assessing reproducibility of clustering patterns observed in analyses of microarray data. *Bioinformatics*, 18(11):1462–1469, 2002.

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