Model order selection for clustered bio-molecular data

Alberto Bertoni, Giorgio Valentini

DSI - Dipartimento di Scienze dell'Informazione,
Università degli Studi di Milano, 20135 Milano, Italy,
{bertoni,valentini}@dsi.unimi.it

1 Introduction

Cluster analysis has been widely applied for investigating structure in bio-molecular data: for instance, 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 [1]. Unfortunately, clustering algorithms may find structure in the data, even when no structure is present instead. Hence we need methods for assessing the validity of the discovered clusters to test the existence of biologically meaningful clusters.

Recently, several methods based on the concept of stability have been proposed to estimate the "optimal" number of clusters in complex bio-molecular data [2, 3]. In this conceptual framework multiple clusterings are obtained by introducing perturbations into the original data, and a clustering is considered reliable if it is approximately maintained across multiple perturbations. In particular, Ben-Hur, Ellisseeff and Guyon proposed to perturb the original data through subsampling procedures, applying then a suitable clustering algorithm to the subsampled data; after estimating the stability of the obtained solutions through a pairwise clustering similarity measure, they assessed the "optimal" number of clusters by means of a visual inspection of the similarity measures across different numbers of clusters [4].

In this paper we propose an improvement of the Ben-Hur algorithm to assess the significance level of the solutions, by introducing a quantitative approach and a statistical test based on the distribution of suitable similarity measures between pairs of clustered projected data. Moreover we propose also a new way to perturb the data, based on random projections into lower dimensional subspaces, that seems to be well-suited to the characteristics (high-dimensionality, redundancy, noise) of genomic and proteomic data.

2 Model Order Selection by Randomized Maps

The proposed *MOSRAM* (Model Order Selection by RAndomized Maps) algorithm perturbs the original data using a randomized mapping $\mu : \mathbb{R}^d \rightarrow \mathbb{R}^{d'}$, with $d' < d$. A key problem with high dimensional genomic data consists in finding a d' such that for every pair $p, q \in \mathbb{R}^d$, the distances between points in the embedded and original metric space are approximately preserved:

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

In [5] we proposed to choose d' according to the *Johnson-Lindenstrauss (JL) lemma* [6]: Given a data set D with $|D| = n$ there exists a $1+\epsilon$-distortion embedding into $\mathbb{R}^{d'}$ with $d' = c \ \log n/\epsilon^2$, where c is a suitable constant. As a consequence, using randomized maps that obey the JL lemma, we may perturb the data introducing only bounded distortions, approximately preserving the structure of the original data [5]. The *MOSRAM* algorithm can be summarized as follows:

MOSRAM algorithm:

```
Input:
    $D$ : a dataset; $k_{max}$: max number of clusters; $n$ : number of pairs of random
    projections; $\mu$ a randomized map; $\mathcal{C}$: a clustering algorithm;
    $sim$ : a clustering similarity measure.
Output:
    $M(i,k)$: list of similarity measures for each $k$ $(1 \leq i \leq n, \ 2 \leq k \leq k_{max})$
begin
    for $k := 2$ to $k_{max}$
        for $i := 1$ to $n$
            $proj_a := \mu(D)$
            $proj_b := \mu(D)$
            $C_a := \mathcal{C}(proj_a, k)$
            $C_b := \mathcal{C}(proj_b, k)$
            $M(i,k) := sim(C_a, C_b)$
end.
```

For each number of clusters k, n similarity measurements between two clusterings C_a and C_b, obtained by applying a suitable clustering algorithm $\mathcal{C}$ to the projected data $proj_a$ and $proj_b$, are computed. It is worth noting that we make no assumptions about the shape of the clusters and in principle any clustering algorithm $\mathcal{C}$, randomized map μ, and clustering similarity measure sim may be used. According to [4] a visual inspection of the empirical distributions of the similarity measures stored in M for different values of k may be used to estimate the optimal number of clusters. Anyway this qualitative approach may lead to subjective decisions, especially when the differences betwen the observed distributions are not clearly defined.

To overcome this problem, we propose a quantitative approach that allows us to estimate the significance of the solutions. We define S_k s.t. $0 \leq S_k \leq 1$ as a random variable that represents the similarity between two k-clusterings (e.g. Fowlkes and Mallows similarity). If S_k values are close to 1 the two clusterings are very similar, if close to 0 very dissimilar. Let be $f(S_k)$ its density function, and

$$F(\bar{s}_k) = \int_{-\infty}^{\bar{s}_k} f(S_k) dS_k \tag{2}$$

its cumulative distribution function. We define $g(k)$ as the integral of the cumulative distribution function:

$$g(k) = \int_0^1 F(S_k) dS_k \tag{3}$$

Intuitively $g(k)$ represents the "concentration" of the similarity values close to 1; that is, if $g(k) \simeq 0$ then the distribution of the values of S_k is concentrated near 1, or, in other words, the k-clustering is stable. On the other hand, if $g(k) \simeq 1$ then the clusterings are totally unstable, while if the distribution is close to the uniform distribution, we have $g(k) \simeq 1/2$. We may directly estimate eq. 3 by numerical integration, or we may more easily obtain $g(k)$ from the estimate of the expectation $E[S_k]$:

$$E[S_k] = \int_0^1 S_k F'(S_k) dS_k = 1 - \int_0^1 F(S_k) dS_k \tag{4}$$

Hence from eq. 4 we may easily compute $g(k)$:

$$g(k) = \int_0^1 F(S_k) dS_k = 1 - E[S_k] \tag{5}$$

Eq. 5 shows also more clearly that we have a very stable and reliable clustering ($E[S_k]$ close to 1), if and only if $g(k)$ is close to 0.

Consider a set of k-clusterings $k \in \mathcal{K}$, where $\mathcal{K}$ is a set of numbers of clusters and the corresponding set of integrals computed according to eq. 3. Then we obtain a set of values $\mathcal{G} = \{g_k | k \in \mathcal{K}\}$. We can sort G obtaining $\hat{\mathcal{G}}$ with values $\hat{g}_i$ in ascending order. For each k-clustering we consider two groups of pairwise clustering similarities values separated by a threshold t^o [1]. Thus we may obtain: $P(S_k > t^o) = 1 - F(S_k = t^o)$, where $F(S_k = t^o)$ is computed according to eq. 2. If n represents the number of trials for estimating the value of S_k then $x_k = P(S_k > t^o)n$ is the number of times for which the similarity values are larger than t^o. The x_k may be interpreted as the successes from $|\mathcal{K}|$ binomial populations with parameters θ_k. If the number of trials n is sufficiently large, and setting X_k as a random variable that counts how many times $S_k > t^o$, we have that the random variables

$$\frac{X_k - n\theta_k}{\sqrt{n\theta_k(1-\theta_k)}} \sim N(0,1) \quad \text{and hence} \quad \sum_{k \in \mathcal{K}} \frac{(X_k - n\theta_k)^2}{n\theta_k(1-\theta_k)} \sim \chi^2$$

Considering the null hypothesis H_0: all the θ_k are equal to θ, where the unknown θ is estimated through its pooled estimate $\hat{\theta} = \frac{\sum_{k \in \mathcal{K}} x_k}{|\mathcal{K}| \cdot n}$, then the null hypothesis may be evaluated against the alternative hypothesis that the θ_k are not all equal using the statistic

$$Y = \sum_{k \in \mathcal{K}} \frac{(x_k - n\hat{\theta})^2}{n\hat{\theta}(1-\hat{\theta})} \sim \chi^2_{|\mathcal{K}|-1} \tag{6}$$

[1] In our experiments we set $t^o = 0.9$.

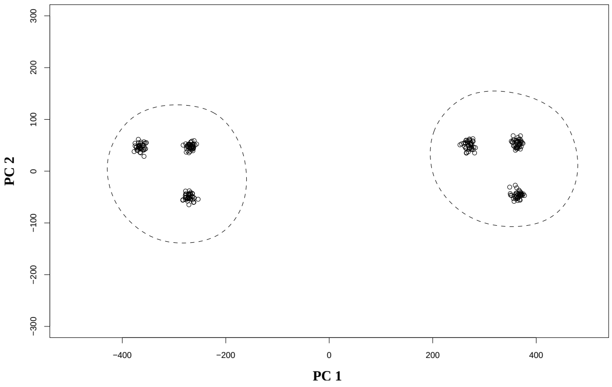

Fig. 1. Synthetic sample: data projected into the two components with highest variance, by means of Principal Component Analysis.

If $Y \geq \chi^2_{\alpha,|\mathcal{K}|-1}$ we may reject the null hypothesis at α significance level, that is we may conclude that with probability $1-\alpha$ the considered proportions are different, and hence that at least one k-clustering significantly differ from the others. Using the above test we start considering all the k-clustering. If a significant difference is registered according to the statistical test we exclude the last clustering (according to the sorting of $\mathcal{G}$). This is repeated until no significant difference is detected: the set of the remaining (top sorted) k-clusterings represent the set of the estimate stable number of clusters discovered (at α significance level).

It is worth noting that the above χ^2-based procedure may be also applied to automatically find the optimal number of clusters using the the similarity measures generated according to the cited Ben-Hur et al. algorithm.

3 Experiments with synthetic and gene expression data

We performed experiments with high dimensional synthetic and real gene expression data to assess the effectiveness of the proposed method.

We considered different high dimensional synthetic data sets, and in any case the correct number of clusters has been detected. To show the ability of our method to discover multiple structures simultaneously present in the data, we propose an experiment with a 1000-dimensional synthetic multivariate gaussian data set, characterized by a two-level hierarchical structure, highlighted by the projection of the data into the two main principal components (Fig. 2): indeed a two-level structure, with respectively 2 and 6 clusters is self-evident in the data. Accordingly, two clusterings (using the Prediction Around Medoid algorithm) are detected at 0.001 significance level by our proposed method, using Plus-Minus-One random projections [5] from a 1000 to a 479-dimensional subspace ($\epsilon = 0.2$). Indeed Fig. 2 shows that the histograms of the similarity measures for

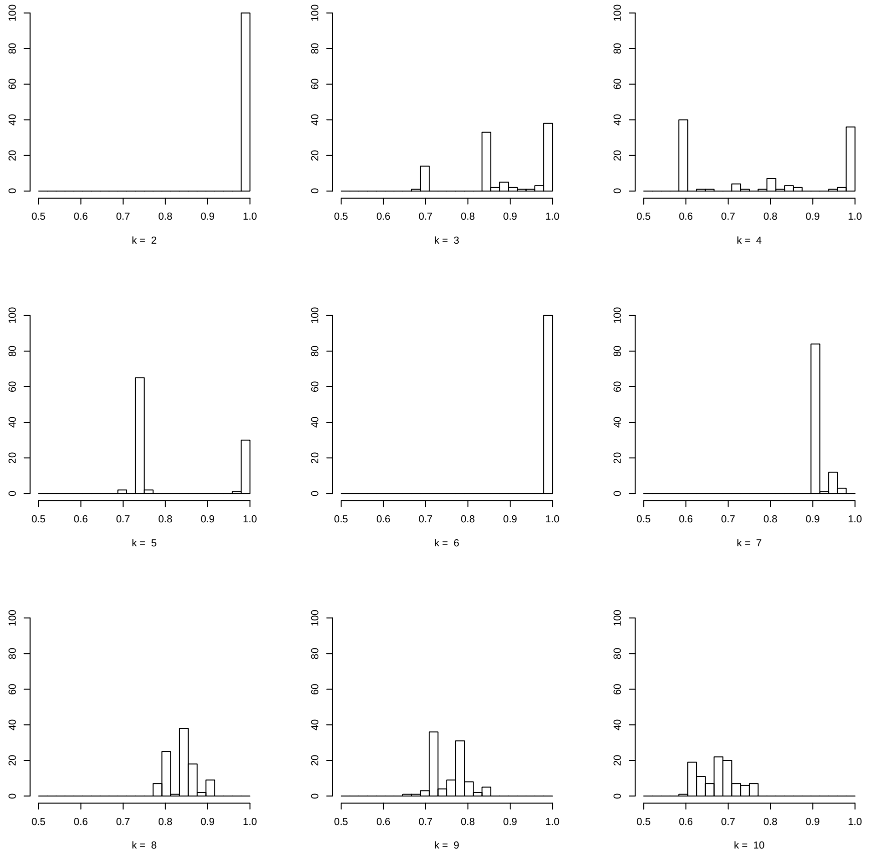

Fig. 2. 1000-dimensional synthetic sample: histograms of the similarity measure distributions for different numbers of clusters.

$k = 2$ and $k = 6$ clusters are tightly concentrated near 1, while for other values of k the similarity measures are spread across multiple values.

Fig. 3 shows the empirical cumulative distribution functions for different numbers of clusters with the *Leukemia* data set [7]. For these experiments we used the classical c-mean clustering algorithm and Plus-Minus-One random projections [5] with $\epsilon = 0.1$ (corresponding to projections from 3574 to 1711-dimensional subspaces, according to the JL lemma). Fig. 3 shows that clusterings with $k = 2$ and $k = 3$ clusters have the smaller area under the empirical cumulative distribution of the similarity values (that is the estimated lowest value of $g(k)$, eq. 3). Our proposed χ^2-based test predicted 2 clusters as the most reliable, with $\alpha = 10^{-5}$ significance level, and the corresponding integral $g(2) = 0.1125$ (eq. 3). Anyway at $\alpha = 10^{-12}$ significance level also 3 clusters are judged quite reliable ($g(3) = 0.2328$). The other numbers of clusters are considered highly

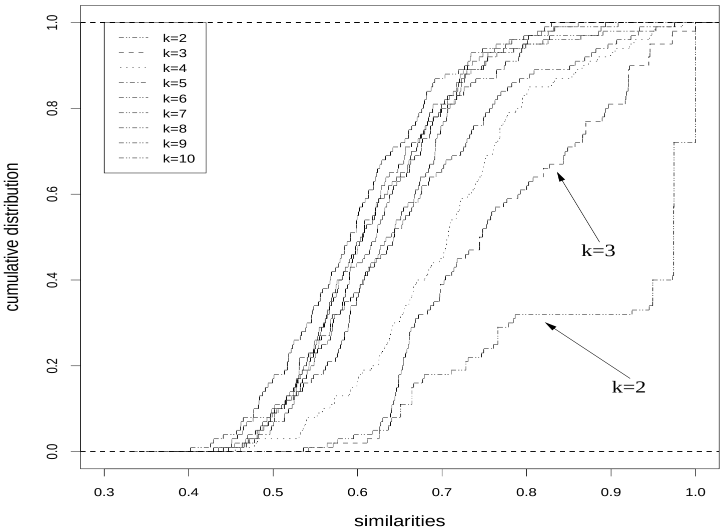

Fig. 3. Leukemia: empirical cumulative distribution functions of the similarity measures for different number of clusters k, $2 \leq k \leq 10$.

unreliable. This perfectly match with the fact that two biologically meaningful groups (ALL, acute lymphoblastic leukemia and AML, acute myeloid leukemia) are present in the data; furthermore ALL can be subdivided into B-cell and T-cell ALL obtaining in this way 3 distinct clusters. Experiments with other DNA microarray data sets confirmed the effectiveness of our approach.

Acknowledgments

This work has been partially funded by the European Pascal Network of Excellence.

References

1. Lapointe, J. et al: PNAS **101** (2004) 811–816
2. Monti, S. et al.: Machine Learning **52** (2003) 91–118
3. Lange, T. et al.: Neural Computation **16** (2004) 1299–1323
4. Ben-Hur, A. et al.: Pacific Symposium on Biocomputing. Vol. 7., (2002) 6–17
5. Bertoni, A., Valentini, G.: Artificial Intelligence in Medicine **37** (2006) 85-109.
6. Johnson, W., Lindenstrauss, J.: In: Vol. 26 of Contemporary Mathematics. (1984)
7. Golub, T.R. et al.: Science **286** (1999) 531–537