Discovering significant structures in clustered data through Bernstein inequality

Alberto Bertoni, Giorgio Valentini

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

Abstract. The reliability of clusters discovered by a given clustering algorithm may be estimated by means of methods based on the concept of stability with respect to "random perturbations" of the data. In this context, a major problem is to estimate the confidence of the measures of reliability; recently proposed procedures realizing this task are correct under the assumption that some probability distributions are normal. Here we discuss a partially "distribution independent" method to assess the statistical significance of the discovered clusterings. Preliminary numerical experiments show the effectiveness of the proposed approach.

1 Introduction

Many clustering algorithms require to "a priori" know the number of clusters to be correctly applied to the unsupervised analysis of the data. In this context, several methods based on the concept of stability have been proposed to estimate the "optimal" number of clusters in clustered data [1, 2]: 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.

Different procedures have been proposed to randomly perturb the data, ranging from bootstrapping techniques [1], to noise injection into the data [3] or random projections into lower dimensional subspaces [4].

For instance, Ben-Hur, Ellisseeff and Guyon proposed to perturb the original data through subsampling procedures, applying then a suitable clustering algorithm to multiple instances of 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 [5].

Even if several works showed that this general approach is effective in discovering structures in complex data [2, 4], a major problem is to estimate the statistical significance of the structures discovered by clustering algorithms. In [6], it is proposed a χ^2-based statistical test of hypothesis to assess the significance of the "optimal" number of clusters: however, some assumptions about the distribution of the similarity measures are needed to estimate the reliability of the obtained clusterings.

In this paper we propose a new distribution-free approach that does not assume any "a priori" distribution of the similarity measures. In the next section we summarize the problem of the assessment of the reliability of a clustering procedure using a stability-based approach. Then we present the method based on Bernstein inequality [7] to assess the statistical significance of the clusterings discovered by a given clustering algorithm, and in the last section we present some numerical experiments to show the effectiveness of the proposed approach.

2 Model order selection through stability based procedures

Let be $\mathcal{C}$ a clustering algorithm, $\rho(D)$ a given random perturbation procedure applied to a data set D and sim a suitable similarity measure between two clusterings (e.g. the Fowlkes and Mallows similarity). For instance ρ may be a random projection from a high dimensional to a low dimensional subspace [8], or a bootstrap procedure to sample a random subset of data from the original data set D [5]; for the similarity measure between two clusterings we may use, for instance, the Jaccard or the Fowlkes and Mallows coefficient [9].

We define S_k ($0 \leq S_k \leq 1$) as the random variable given by the similarity between two k-clusterings obtained by applying a clustering algorithm $\mathcal{C}$ to pairs D_1 and D_2 of random independently perturbed data. The intuitive idea is that if S_k is concentrated close to 1, the corresponding clustering is stable with respect to a given controlled perturbation and hence it is reliable.

Let be $f_k(s)$ its density function, and

$$F_k(\bar{s}) = \int_{-\infty}^{\bar{s}} f_k(s)ds \tag{1}$$

its cumulative distribution function.

A parameter of concentration implicitly used in [5] is the integral $g(k)$ of the cumulative distribution F_k:

$$g(k) = \int_0^1 F_k(s)ds \tag{2}$$

We can observe the following facts:
Fact 1: $g(k)$ is strictly related to the expectation of S_k; indeed:

$$E[S_k] = \int_0^1 s f_k(s)ds = \int_0^1 s F_k'(s)ds = 1 - \int_0^1 F_k(s)ds = 1 - g(k) \tag{3}$$

Fact 2: if $g(k) \simeq 0$, S_k in concentrated close to 1; indeed, from $0 \leq S_k \leq 1$ it follows $S_k^2 \leq S_k$; hence:

$$Var[S_k] = E[S_k^2] - E[S_k]^2 \leq E[S_k] - E[S_k]^2 = g(k)(1 - g(k)) \tag{4}$$

Hence, $g(k) \simeq 0$ implies $Var[S_k] \simeq 0$.

In conclusion, considerings Facts 1 and 2, $g(k)$ or equivalently $E[S_k]$ can be used as a good index of the reliability of the k-clusterings (clusterings with k clusters).

For every k ($2 \leq k \leq H+1$), we can estimate $E[S_k]$ by means of multiple similarity measures S_{kj}:

$$S_{kj} = sim\left(\mathcal{C}(\rho_{kj}^{(1)}(D), k), \mathcal{C}(\rho_{kj}^{(2)}(D), k)\right), \quad 1 \leq j \leq n \tag{5}$$

S_{kj} represents the similarity between two k-clusterings obtained through the application of the algorithm $\mathcal{C}$ to the perturbed data $\rho_{kj}^{(1)}(D)$ and $\rho_{kj}^{(2)}(D)$.

$E[S_k]$ may be estimated by the empirical means ξ_k:

$$\xi_k = \sum_{j=1}^{n} \frac{S_{kj}}{n} \tag{6}$$

We may perform a sorting of the ξ_k:

$$(\xi_2, \xi_3, \ldots, \xi_{H+1}) \stackrel{sort}{\rightarrow} (\xi_{p(1)}, \xi_{p(2)}, \ldots, \xi_{p(H)}) \tag{7}$$

where p is the index permutation such that $\xi_{p(1)} \geq \xi_{p(2)} \geq \ldots \geq \xi_{p(H)}$.

Exploiting the ordering of the empirical means that represents the most reliable $p(1)$-clustering down to the least reliable $p(H)$-clustering, we would establish which are the significant clusterings (if any) discovered in the data, considering that eq.3 represents the "goodness" of the clustering in terms of its stability.

To this end we would estimate if for a given r, $2 \leq r \leq H$, there exists a statistically significant difference between the reliability of the best $p(1)$ clustering and the $p(r)$ clustering. In other words we may state the null hypothesis H_0 and the alternative hypothesis in the following way:

H_0**:** $p(1)$ clustering is not more reliable than $p(r)$ clustering, that is $E[S_{p(1)}] \leq E[S_{p(r)}]$
H_a**:** $p(1)$ clustering is more reliable than $p(r)$ clustering, that is $E[S_{p(1)}] > E[S_{p(r)}]$

3 Hypothesis testing based on Bernstein inequality

We briefly recall the Bernstein inequality, because this inequality is used to build-up our proposed hypothesis testing procedure.

Bernstein inequality. If $Y_1, Y_2, \ldots, Y_n$ are independent random variables s.t. $0 \leq Y_i \leq 1$, with $\mu = E[Y_i], \sigma^2 = Var[Y_i], \bar{Y} = \sum Y_i/n$ then

$$Prob\{\bar{Y} - \mu > \Delta\} \leq e^{\frac{-n\Delta^2}{2\sigma^2 + 2/3\Delta}} \tag{8}$$

Our goal is to apply the Bernstein inequality to assess the significance of a given k-clustering with respect to another k'-clustering, $k \neq k'$, exploiting the ordering of the computed empirical means of the similarity measures (eq. 7).

Consider the following random variables:

$$P_i = S_{p(1)} - S_{p(i)} \quad \text{and} \quad X_i = \xi_{p(1)} - \xi_{p(i)} \tag{9}$$

We start considering the first and last ranked clustering $p(1)$ and $p(H)$. In this case the above null hypothesis H_0 becomes: $E[S_{p(1)}] \leq E[S_{p(H)}]$, that is: $E[S_{p(1)}] - E[S_{p(H)}] = E[P_H] \leq 0$. The distribution of the random variable X_H (eq. 9) is in general unknown; anyway note that in the Bernstein inequality no assumption is made about the distribution of the random variables Y_i (eq. 8). Hence, fixing a parameter $\Delta \geq 0$, considering true the null hypothesis $E[P_H] \leq 0$, and using Bernstein inequality, we have:

$$Prob\{X_H \geq \Delta\} \leq Prob\{X_H - E[P_H] \geq \Delta\} \leq e^{\frac{-n\Delta^2}{2\sigma^2 + 2/3\Delta}} \tag{10}$$

Considering an instance (a measured value) $\hat{X}_H$ of the random variable X_H, if we let $\Delta = \hat{X}_H$ we obtain the following probability of type I error:

$$P_{err}\{X_H \geq \hat{X}_H\} \leq e^{\frac{-n\hat{X}_H^2}{2\sigma_H^2 + 2/3\hat{X}_H}}$$

with $\sigma_H^2 = \sigma_{p(1)}^2 + \sigma_{p(H)}^2$.

If $P_{err}\{X_H \geq \hat{X}_H\} < \alpha$, we reject the null hypothesis: a significant difference between the two clusterings is detected at α significance level and we continue by testing

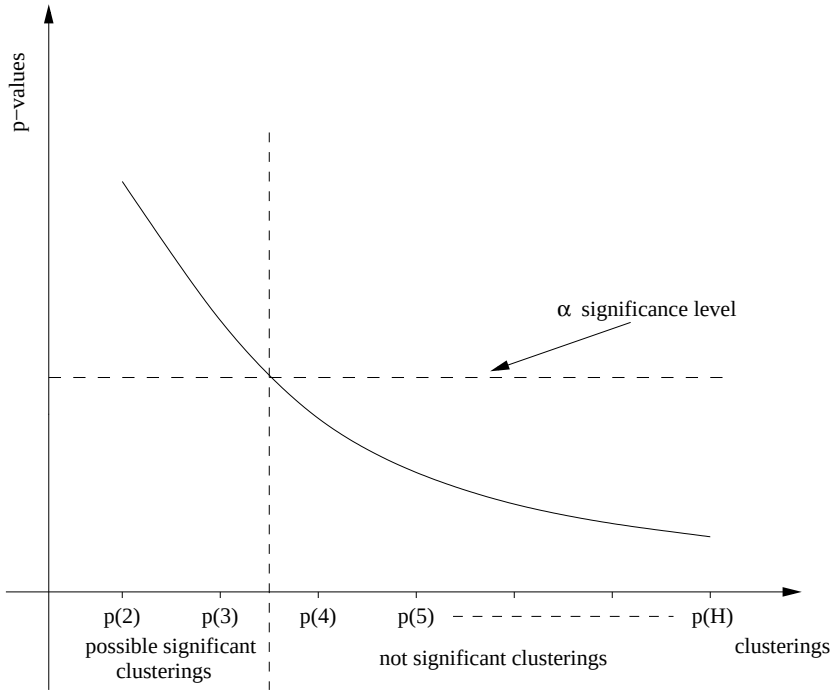

Figure 1: P-values computed for different number of clusters. Possibly significant and not significant clusterings detected at α-significance level.

the $p(H-1)$ clustering. More in general if the null hypothesis has been rejected for the $p(H-r+1)$ clustering, $1 \leq r \leq H-2$ then we consider the $p(H-r)$ clustering, and by union bound we can estimate the type I error:

$$P_{err}(H-r) = Prob\{\bigvee_{H-r \leq i \leq H} X_i \geq \hat{X}_i\} \leq \sum_{i=H-r}^{H} Prob\{X_i \geq \hat{X}_i\} \leq \sum_{i=H-r}^{H} e^{\frac{-n\hat{X}_i^2}{2\sigma_i^2+2/3\hat{X}_i}} \tag{11}$$

As in the previous case, if $P_{err}(H-r) < \alpha$ we reject the null hypothesis: a significant difference is detected between the reliability of the $p(1)$ and $p(H-r)$ clustering and we iteratively continue the procedure estimating $P_{err}(H-r-1)$.

This procedure stops if either of these cases succeeds:
I) The null hypothesis is rejected till to $r = H-2$, that is $\forall r, 1 \leq r \leq H-2, \quad P_{err}(H-r) < \alpha$: in this case all the possible hypotheses have been rejected and the only reliable clustering at α-significance level is the top ranked one, that is the $p(1)$ clustering.
II) The null hypothesis cannot be rejected for $r < H-2$, that is, $\exists r, 1 \leq r \leq H-2, \quad P_{err}(H-r) \geq \alpha$: in this case the clusterings that are significantly less reliable than the top ranked $p(1)$ clustering are the $p(r+1), p(r+2), \ldots, p(H)$ clusterings.

Note that in this second case we cannot state that there is no significant difference between the first r top-ranked clusterings, since the upper bound provided by the Bernstein inequality is not guaranteed to be tight. This situation is depicted in Fig. 1. In this case, for a given α-significance level, clusterings from $p(4)$ to $p(H)$ are significantly less reliable than the top ranked clustering, but we cannot say anything about the reliability of $p(2)$ and $p(3)$ clusters.

To answer to this question, we may apply the χ^2-based hypothesis testing proposed in [6] to the remaining top ranked clusterings to establish which of them are significant at α level, but in this case we need to assume that the similarity measures between

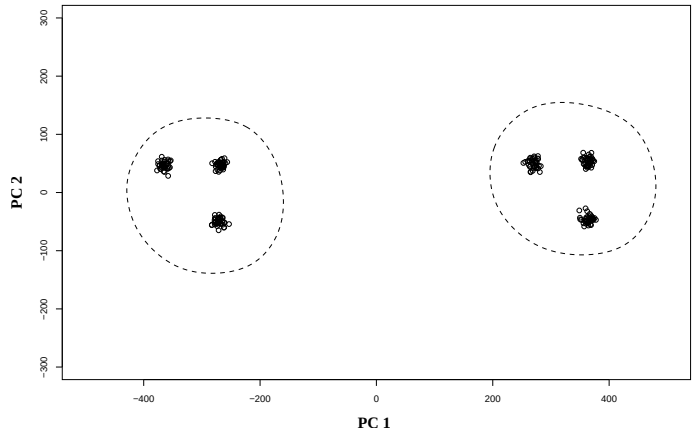

Figure 2: Synthetic sample: data projected into the two components with highest variance, by means of Principal Component Analysis.

pairs of clusterings are distributed according to a normal distribution. For instance, in the case of Fig. 1, we could apply the the χ^2-based test to the clusterings $p(1), p(2)$ and $p(3)$).

4 Numerical experiments

We present some preliminary experiments with high dimensional synthetic data and gene expression data to assess the effectiveness of the proposed method. In particular 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. To perturb the data we applied Plus-Minus-One random projections [4] from a 1000 to a 479-dimensional subspace using the *clusterv* R library [10] and developing a new R package (*MOSCLUST*, Model Order Selection for CLUSTering) for implementing the Bernstein-based hypothesis testing procedure. We applied the Prediction Around Medoids clustering algorithm [11], considering the reliability of clusterings composed from 2 to 10 clusters. The p-values computed according to our hypothesis testing procedure based on Bernstein inequality are summarized in Table 1.

In Tab. 1 the clusterings are sorted according to the value of the empirical means ξ of the similarity measures computed according to eq. 6. Accordingly to the known

Table 1: Synthetic data set. P-values computed according to the Bernstein inequality for different numbers of clusters. ξ are empirical means of the similarity measures computed according to eq. 6. The clusterings are sorted according to ξ.

Num.clusters	**p-values**	ξ
2	———-	1.0000
6	1.0000e+00	1.0000
7	3.5932e-06	0.9217
8	7.8628e-10	0.8711
9	8.7535e-14	0.8132
5	1.4194e-15	0.8090
3	1.0828e-16	0.8072
10	8.5029e-17	0.7715
4	5.6677e-20	0.7642

Table 2: Leukemia data set. P-values computed according to the Bernstein inequality for different numbers of clusters. ξ are empirical means of the similarity measures computed according to eq. 6. The clusterings are sorted according to ξ.

Num.clusters	**p-values**	ξ
2	———-	0.8874
3	2.1418e-07	0.7671
4	4.7712e-11	0.7078
5	2.5084e-14	0.6607
7	4.6057e-15	0.6352
6	1.4496e-15	0.6289
8	6.5054e-16	0.6185
10	3.0212e-16	0.6162
9	1.2274e-17	0.5968

characteristics of these data, with a significance level $\alpha = 0.01$ all the clusterings except the clustering with 6 clusters are considered significantly less reliable than the first top ranked clustering (indeed their p-values are all largely below 0.01). Hence only the clustering with 6 clusters is considered potentially reliable as the top ranked one (a 2-clustering).

As a second example we propose an experiment with DNA microarray data, using the publicly available *Leukemia* data set [12]. For these experiments we filtered the genes according to the procedures described in [12]. We used the classical c-mean clustering algorithm and Plus-Minus-One random projections from the original space of 3574 gene expression levels to 1711-dimensional subspaces according according to a bounded data distortion of about 10% [4].

Tab. 2 shows the p-values of the clusterings sorted according to their ξ values (eq. 6). Our proposed procedure detects the $2 - clustering$ as the most reliable at 0.01 level, according to the fact that two biologically meaningful groups (ALL, acute lymphoblastic leukemia and AML, acute myeloid leukemia) are present in the data. Choosing a significance level $\alpha = 10^{-7}$ we cannot reject the null hypothesis that a 2-clustering is less or equally reliable than a 3-clustering: indeed ALL can be subdivided into B-cell and T-cell ALL, obtaining in this case 3 classes.

5 Conclusion

We proposed a test of hypothesis based on Bernstein inequality to estimate if there is a significant difference between the reliability of two clusterings performed on the same data. It does not assume that the similarity measures used to estimate the reliability of the clusterings are distributed according to a normal or any other distribution. This testing procedure may be applied to any stability-based procedure to assess the reliability of the clusterings, using random projections, bootstrapping techniques or noise injection procedures to perturb the original data.

Acknowledgments

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. Monti, S., Tamayo, P., Mesirov, J., Golub, T.: Consensus Clustering: A Resampling-based Method for Class Discovery and Visualization of Gene Expression Microarray Data. Machine Learning **52** (2003) 91–118
2. Lange, T., Roth, V., Braun, M., Buhmann, J.: Stability-based validation of clustering solutions. Neural Computation **16** (2004) 1299–1323
3. McShane, L., Radmacher, D., Freidlin, B., Yu, R., Li, M., Simon, R.: Method for assessing reproducibility of clustering patterns observed in analyses of microarray data. Bioinformatics **18** (2002) 1462–1469
4. Bertoni, A., Valentini, G.: Randomized maps for assessing the reliability of patients clusters in DNA microarray data analyses. Artificial Intelligence in Medicine **37** (2006) 85–109
5. Ben-Hur, A., Ellisseeff, A., Guyon, I.: A stability based method for discovering structure in clustered data. In Altman, R., Dunker, A., Hunter, L., Klein, T., Lauderdale, K., eds.: Pacific Symposium on Biocomputing. Volume 7., Lihue, Hawaii, USA, World Scientific (2002) 6–17
6. Bertoni, A., Valentini, G.: Model order selection for clustered bio-molecular data. In: Workshop on Probabilistic Modeling and Machine Learning in Structural and Systems Biology, Tuusula, Finland (2006)
7. Hoeffding, W.: Probability inequalities for sums of independent random variables. J. Amer. Statist. Assoc. **58** (1963) 13–30
8. Achlioptas, D.: Database-friendly random projections. In Buneman, P., ed.: Proc. ACM Symp. on the Principles of Database Systems. Contemporary Mathematics, New York, NY, USA, ACM Press (2001) 274–281
9. Jain, A., Dubes, R.: Algorithms for clustering data. Prentice Hall, Englewood Cliffs, NJ (1988)
10. Valentini, G.: Clusterv: a tool for assessing the reliability of clusters discovered in DNA microarray data. Bioinformatics **22** (2006) 369–370
11. Kaufman, L., Rousseeuw, P.: Finding Groups in Data: An Introduction to Cluster Analysis. Wiley, New York (1990)
12. Golub, T., et al.: Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. Science **286** (1999) 531–537