

Bagged Biclustering for Microarray Data

Blaise Hanczar¹ Mohamed Nadif²

Abstract. One of the major tools of transcriptomics is the biclustering that simultaneously constructs a partition of both examples and genes. Several methods have been proposed for microarray data analysis that enables to identify groups of genes with similar expression profiles only under a subset of examples. We propose to improve the quality of these biclustering methods by using an ensemble approach. Our bagged biclustering method generates a collection of biclusters using the bootstrap samples of the original data and aggregate them into new biclusters. Our method improve the performance of biclustering on artificial and real datasets.

1 Introduction

The capacity of microarray to measure simultaneously the expression of a whole genome under different experimental condition, is of great interest for biologists. Biclustering methods allow the identification of relevant groups of genes and conditions that cannot be identified by classic clustering techniques. These kinds of methods consist in simultaneous clustering on rows and columns, to reorganize the data set into homogeneous blocks. Several biclustering algorithms have been proposed and used on microarray data [1, 4, 6]. In this paper, we try to improve the performance of biclustering algorithms in using the ensemble approach. The principle of ensemble methods is to construct a set of models, then to aggregate them into a single model, generally by using generally a voting scheme. In this paper, we propose in this paper a bagging approach for the biclustering of microarray data. Although the problem of ensemble biclustering shares some traits with classic biclustering, there are two major issues which are specific to ensemble methods. The first one is the generation of a collection of biclusters. This raises the question of how to generate different biclusters and of the source of diversity ? We have chosen the bagging approach to generate biclusters from bootstrapped datasets. The second issue is about the aggregation function. How to combine the different biclusters and resolve the label correspondence problem? Based on the works of Sterhl and Ghosh [7], we propose a solution that consists in creating *meta-clusters* of biclusters. Finally, for each gene and example, we compute the probabilities of their belonging to each meta-cluster. We test our method both on artificial and real data. The results on artificial data show that ensemble methods enable to improve the accuracy of biclustering. The results obtained on real data show that ensemble biclusters give a lower residue than classic biclusters. Moreover the ensemble biclusters are also biologically more relevant with respect to the prior knowledge of data.

2 Methods

The principle of bagged biclustering consists in applying a biclustering method on multiple bootstrapped datasets and aggregate the results. Our method can be divided into 3 steps: construction of a collection of biclusters, identification of the meta-cluster, computation of the biclusters. In the first step we generate a high number of different biclusters. To do so, we generate bootstrap samples of the original data. We apply the bootstrap sampling only on the genes. On each of the R bootstrapped datasets, a biclustering algorithm, with the same parameters, is applied to produce K biclusters. We obtain a collection of KR biclusters noted B^b that are used to identify meta-clusters. The aim of the second step is to identify K meta-clusters merging the similar biclusters. The idea is that if two biclusters, generated from different bootstrapped data, are similar, it is likely that they represent the same bicluster. All bootstrapped biclusters representing the same bicluster should be grouped into a meta-cluster. The notion of similarity between two biclusters depends on the number of elements (genes and examples) they have in common. Here we use the Jaccard index to evaluate this similarity. From this similarity we compute a distance matrix and a hierarchical clustering of the biclusters using the average linkage. From the obtained dendrogram we can identify K meta-clusters in cutting the dendrogram. Before deducing these meta-clusters in the third step, we estimate the probability of each gene and example to belong to the meta-clusters. The last step consists in computing the final biclusters of the original data. Each meta-cluster is assigned to a bicluster. Then each element can be assigned to biclusters depending on computed probabilities. We define a threshold t , if the probability is higher than t then the gene or example is associated to the bicluster.

3 Results and discussion

In our simulation study, we evaluate the performance of bagged biclustering and compare it to single biclustering. We performed our experiments on artificial and real datasets with the five algorithms: Bimax, Cheng & Church, plaid model, spectral biclustering and Xmotifs.

Figure 1 shows the biclustering error on artificial data including 4 biclusters. In each graphic, the error biclustering is computed in function of the quantity of noise introduced in the dataset. The dot curve represents the error of the single biclustering algorithm, the triangle curve the bagged biclustering. We see that in all cases the error is naturally increasing with the noise. For Bimax, we see that at $\sigma = 0.4$ the error of single biclustering "jumps" from 0 to 1700. The error of the bagged biclustering increases slowly to join the error of the single biclustering at $\sigma = 1.8$. We can interpret this as follows: for $\sigma \leq 0.4$ the biclustering problem is easy, the performance of biclustering is maximal. For $\sigma \geq 1.8$ the datasets are so noisy that all biclusterings are meaningless, we have checked that the error is the

¹ LIPADE, University Paris Descartes

² LIPADE, University Paris Descartes

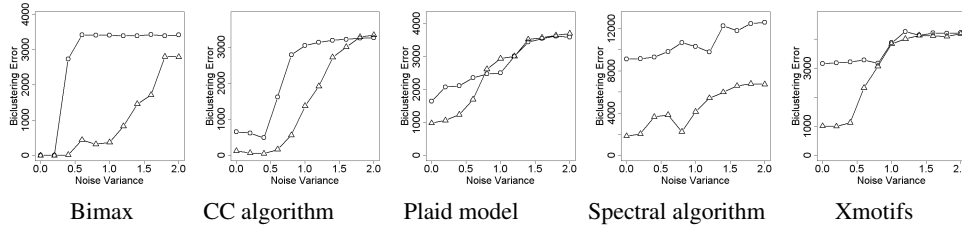


Figure 1. Biclustering error in function of noise variance on artificial data when $K = 4$. The two graphic lines correspond respectively to results with Bimax, CC algorithm, plaid model, Spectral biclustering and Xmotifs. Dot lines and triangle lines correspond respectively to single and bagged biclustering.

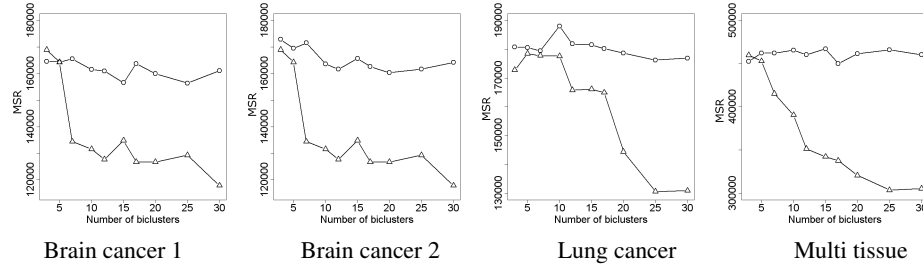


Figure 2. Application of the CC algorithm on each dataset (Brain cancer 1, Brain cancer 2, Lung cancer and Multi tissue). Each column corresponds to a dataset, the first line to the plaid model and second line to the CC algorithm. For each dataset, MSR is represented in function of the number of biclusters K . Dot lines and triangle lines correspond respectively to MSR of single and bagged biclustering.

same as a random biclustering. These graphics show that bagged Bimax gives a better biclustering than single Bimax. We find the same type of results in the major part of our experiments. The Cheng & Church algorithm and Bimax have the same behavior. Error of single CC increases faster than bagged CC. In plaid model results, we see that the error of the bagged plaid model is lower than that of the single plaid model for $\sigma < 1$. For higher level of noise, $\sigma \geq 1$, the two algorithms give the same performance as "random" biclustering. The results of Spectral biclustering are singular since the error does not jump for a given value of σ but is regularly increasing with σ . The error curves of single and bagged spectral biclustering are almost parallel. The error of bagged spectral biclustering is always below the error of the single spectral biclustering. For Xmotifs results, the error curve of bagged Xmotifs always begins below the curve of single Xmotifs, then joins it for higher levels of noise. The results, including various biclustering algorithms and different number of biclusters (not reported here), show that bagged biclustering gives better results than single biclustering.

We applied single and bagged biclustering with the plaid model and the CC models on four microarray datasets. We computed the MSR of results produced by each method according different values of K . The number of bootstrap iterations for bagged biclustering was $R = 100$. Figure 2 which consists of eight plots, shows the MSR of the single (dot lines) and bagged (triangle lines) biclustering in function of the number of biclusters K . The results with the plaid model exhibit a lot of variability, but the MSR of bagged biclustering is lower than single biclustering. In the four datasets there is a peak phenomenon, MSR is firstly decreasing then increasing with the number of biclusters. The results with CC algorithm are clearer and we note that the performance increases with the number of biclusters. The MSR of single biclustering remains stable whereas the MSR of bagged biclustering decreases strongly and becomes stable for high number of biclusters. In all datasets, bagged biclustering is

much better than single biclustering for $K > 7$.

4 Conclusion

In this paper we have introduced the concept of ensemble methods for biclustering in the context of microarray data. On artificial data, we have shown that ensemble method enables to strongly decrease the biclustering error compared to classic methods whatever the used biclustering algorithm, the number of biclusters and the chosen noise level. On real data, bagged biclustering provides biclusters more relevant than single biclustering according to their MSR value. The use of our approach is a new powerful tool for microarray analysis and should allow biologists to identify new relevant patterns in gene expression data.

REFERENCES

- [1] Y. Cheng and G. M. Church, 'Biclustering of expression data.', *Proc Int Conf Intell Syst Mol Biol*, **8**, 93–103, (2000).
- [2] Thomas G. Dietterich, 'Ensemble methods in machine learning', *Lecture Notes in Computer Science*, **1857**, 1–15, (2000).
- [3] Y. Kluger, R. Basri, J. T. Chang, and M. Gerstein, 'Spectral biclustering of microarray data: coclustering genes and conditions.', *Genome Res*, **13**(4), 703–716, (2003).
- [4] Laura Lazzaroni and Art Owen, 'Plaid models for gene expression data', Technical report, Stanford University, (2000).
- [5] S. C. Madeira and A. L. Oliveira, 'Biclustering algorithms for biological data analysis: a survey', *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, **1**(1), 24–45, (2004).
- [6] Amela Prelic, Stefan Bleuler, Philip Zimmermann, Anja Wille, Peter Buhlmann, Wilhelm Gruissem, Lars Hennig, Lothar Thiele, and Eckart Zitzler, 'A systematic comparison and evaluation of biclustering methods for gene expression data', *Bioinformatics*, **22**(9), 1122–1129, (2006).
- [7] A. Strehl and J. Ghosh, 'Cluster ensembles - a knowledge reuse framework for combining multiple partitions', *Journal of Machine Learning Research*, **3**, 583–617, (2002).