

A nonparametric algorithm for clustering microarray data

Riccardo De Bin and Davide Risso

Abstract Cluster analysis is a crucial tool in several biological and medical studies dealing with microarray data. Such studies pose challenging statistical problems due to dimensionality issues, being the number of variables much higher than the number of observations. Here, we present a novel approach to clustering of microarray data via nonparametric density estimation, based on the following steps: (i) selection of relevant variables; (ii) dimensionality reduction; (iii) clustering of observations in the reduced space. Applications on simulated and real data show promising results in comparison with those produced by two standard approaches, *k-means* and *Mclust*. In the simulation studies, our nonparametric approach shows performances comparable to those of models based on normality assumption. On the other hand, in two benchmarking real datasets, it outperforms the existing parametric approaches.

Key words: Cluster analysis, Dimensionality reduction, Genomics and Microarray data analysis, Nonparametric/semiparametric statistical methods

1 Introduction

The analysis of gene expression microarray data using clustering techniques plays an important role, for instance, in the discovery, validation, and understanding of various classes and subclasses of cancer. Three main types of statistical problems arise in classification of cancer samples in a microarray experiment [5]: (a) identification of new classes using gene expression profiles (cluster analysis/unsupervised

Riccardo De Bin

Department of Statistical Sciences, University of Padua, via Cesare Battisti 241, 35121 Padova
e-mail: debin@stat.unipd.it

Davide Risso

Department of Statistical Sciences, University of Padua, via Cesare Battisti 241, 35121 Padova
e-mail: davide.risso@unipd.it

learning); (b) classification of samples into known classes (discriminant analysis/supervised learning); (c) identification of “marker” genes which characterize the difference among the classes (variable selection). For a recent review of clustering techniques for gene expression data, see [11].

In recent years, the improvement of computational resources has enabled to pursue new clustering techniques, or to develop ideas putted aside due to computational problems. In this context, Azzalini and Torelli [2] proposed a method, called *pdfClust*, based on a nonparametric estimate of the underlying density function.

In this paper we present a method to handle simultaneously problems (a) and (c) previously mentioned. In particular, we discuss an application of *pdfClust* to microarray data (for details on the method we refer the reader to the original article [2]). We compare it with the traditional algorithms, such as *k-means* algorithm and its direct competitor, *Mclust*, a state-of-the-art mixture-model-based clustering tool [6]. We follow the strategy presented by McLachlan et al. [13], which consists on a data dimensional reduction in order to focus on most significant dimensions. We modify their method by using a nonparametric approach, achieving improvements in clustering of samples both in simulated and in real experiments.

2 A novel algorithm to clustering of expression data

To be consistent with microarray applications, we will denote with “genes” the p variables and with “samples” the n observations. Nonetheless, it should be clear that the approach proposed here is not limited to microarray data, but, in principle, it could be applied to every application with “large p , small n ”.

McLachlan et al. [13] propose a mixture model-based approach to cluster microarray expression data. Their scheme accounts for gene selection, through mixtures of t distributions, and dimensionality reduction, through a mixture of factor analyzers. See [13] for details. Although their approach sounds as a good possibility to cluster samples in a high-dimensional space, there are two main limitations. Firstly, the parametric assumptions about clusters distributions can be restrictive [12]; for example, two Gaussian random variables can result in a single mode (one cluster) or even a two component multivariate Gaussian mixture can lead to more than two modes [12]. Moreover, it needs pre-specification of the number of the mixture components. This, from an unsupervised perspective, which assumes that the true number of clusters is unknown, represents a serious limitation.

Our approach can be summarized as follows: (i) cluster samples p times using the univariate distribution of each gene and select for the subsequent analyses the genes, p' say, that recognize two or more groups in the data; (ii) reduce dimensionality from p' to p'' with some data reduction technique; (iii) apply *pdfClust* algorithm in the p'' -dimensional space.

As for step (i), i.e., *gene selection*, we consider a gene relevant if its values in one category (healthy, say) are different from the ones in the other category (unhealthy, say) or categories; in other words, the samples are in different clusters. In

this way, it seems reasonable to apply a cluster method for each gene, and retain as relevant the ones for which the method recognizes different clusters. In a non-parametric framework, we can apply to each gene *pdfClust*, taking advantage of the self-detection number of clusters feature. We select only the genes for whose the method recognizes two or more clusters.

As for step (ii), i.e., *dimensionality reduction*, considered if the selected genes are still too many, we propose to keep the first principal components, as in [2].

3 Simulation models

For simulating data with structure similar to that of real microarray experiments, we use two schemes, i.e., the Gamma-Gamma model [10] and the Normal-Uniform model [7].

Gamma-Gamma (GG) model

The samples are assumed to be independently generated from Gamma distributions with a constant shape parameter α and gene-specific random scales $\lambda_i, i = 1, \dots, p$; λ_i is assumed to have a Gamma distribution with shape hyperparameter α_0 and scale hyperparameter v . The genes are generated to be either “equally expressed” (i.e. one group) or “differentially expressed” (i.e. two groups) among the samples. We generated $n = 100$ samples and $p = 2,000$ genes, each with probability 0.05 of being differentially expressed. We applied the algorithm stated in Section 2 repeating the procedure $B = 5,000$ times.

Normal-Uniform (NU) model

The model deals with k -class classification of samples, for general k . It is based on a mixture of Normal and Uniform distributions (see [7] for details of the model). We exploit the model to simulate gene expressions for a three-class problem, similar to that of the leukaemia data (see Section 4).

We simulated $B = 5,000$ datasets of $n = 100$ samples, $p = 1,000$ genes and $m = 3$ classes defined as follows: *class 1* consists of 40 samples with 150 up-regulated and 50 down-regulated genes; *class 2* consists of 40 samples with 50 down-regulated genes; *class 3* consists of 20 samples with neither up- nor down-regulated genes. Note that classes 2 and 3 are “close” to each other with respect to class 1.

4 Real data

Along with simulations, we consider two benchmarking real datasets, studied before by several Authors [1, 4, 5, 8, 9, 13], to which we will refer as the colon data and the leukaemia data.

Colon data

Alon et al. [1] used Affymetrix oligonucleotide arrays to measure the expression of 6,500 human genes in 40 tumor and 22 normal colon tissue samples. They focused on the subset of 2,000 genes with highest minimal intensity across the samples: the raw expression values of these 2,000 genes comprise our dataset. Following the notation in [13], we named 1-40 the tumor samples and 41-62 the normal ones.

Leukaemia data

Golub et al. [9] studied the gene expression of two types of acute leukaemias, acute lymphoblastic leukaemia (ALL) and acute myeloid leukaemia (AML). Gene expression levels were measured using Affymetrix oligonucleotide arrays containing 6,817 human genes. The dataset comprises 47 cases of ALL (38 B-cell and 9 T-cell) and 25 cases of AML. The classification of samples is more difficult in this example than in colon data because it is much harder to classify between subclasses of the same plasticity than to distinguish between healthy and cancer tissues. Moreover, we have a typical hierarchical structure, since B-cell and T-cell are subclasses of the ALL class, harder to separate than AML from ALL. We follow the preprocessing steps described in [5]. This procedure left us with 3,892 genes.

5 Results and discussion

5.1 Simulated data

GG model

Table 1 shows that both *pdfClust* and *Mclust* provide results surprisingly good in correct cluster recognition, low error rate and high sensitivity/specificity: this could be explained by an extreme distance between the two groups in the original p -dimensional space.

More interesting is the very different behaviour in the choice of the relevant genes: *pdfClust* is very good in recognizing them, with a very low error rate (about 0.08), while *Mclust* fails in this field, since it selects relevant genes very badly (with error rate about 78%).

NU model

As expected, Table 2 shows that in this model both *pdfClust* and *Mclust* lead to higher classification errors than in GG model. Also in the gene selection step, both methods have difficulties in finding the relevant genes. Nevertheless, *pdfClust* outperforms *Mclust* according to the gene selection error rate (“RG” row in Table 2).

Mclust is able to recognize three clusters in the 39% and two clusters in the 34% of the simulations; *pdfClust* recognizes three clusters in the 19% and two clusters in the 47% of the simulations. On the other hand, the mean error rate of the final classification is 0.135 for *pdfClust* while for *Mclust* is 0.227.

Table 1 Simulation results for *pdfClust* and *Mclust* in GG model: correct number of cluster (CC), sensitivity (SE), specificity (SP), error rate (ER) and error rate in the selection of relevant genes (RG).

	PC		MC	
	mean	se	mean	se
SE	0.9877	0.0007	0.9991	0.0001
SP	0.9866	0.0008	0.9985	0.0004
ER	0.0128	0.0006	0.0012	0.0002
RG	0.0837	0.0061	0.7787	0.0093
CC	0.77		0.84	

Table 2 Simulation results for *pdfClust* and *Mclust* in NU model: rate of two clusters identification (CC2), rate of three clusters identification (CC3), error rate (ER) and error rate in the selection of relevant genes (RG).

	PC		MC	
	mean	se	mean	se
ER	0.135	0.004	0.227	0.005
RG	0.433	0.041	0.616	0.077
CC2	0.47		0.34	
CC3	0.19		0.39	

5.2 Real data

Colon data

As described in Section 2, we analyzed the dataset, following three steps: (i) gene selection, (ii) dimensionality reduction, (iii) clustering of samples. Namely, the first step of the procedure consists in applying the cluster algorithm to the univariate distribution of each gene. The genes that recognize two or more clusters are considered for the further steps.

In the first step, *pdfClust* algorithm was able to recognize 84 genes, which discriminate data in two or more groups. We proceeded by considering the first three principal components of this reduced data-matrix. The procedure found three clusters, summarized in Table 3, which clearly correspond to biologically meaningful groups. The first cluster consists of tumor tissues (with 3 misclassified), while clusters 2 and 3 comprise normal tissues (with 5 misclassified). It is worth noting that six out of the eight misallocated samples (tumor tissues 30, 33 and 36 and normal tissues 48, 58 and 60) are found to be misclassified in several previous analyses, including [4, 13]. As stated for instance in [4], these six samples are likely to be wrongly labeled. Furthermore, Getz et al. [8] reported that there was a change in the protocol during the experiments: tumor samples 1-11 and normal samples 41-51 were collected within the first protocol, while tumor samples 12-40 and normal samples 52-62 were collected within the second. Although for the tumor samples

Table 3 Clusters found after *pdfClust* procedure in Colon data; tumor samples are labeled 1-40, normal samples 41-62; misallocated samples are shown in bold font.

Cluster 1	1-6,8-19,21-29,31,32,34,35,37-40, 48,58,60
Cluster 2	7 ,41-47,49-52
Cluster 3	20,30,33,36 ,53-57,59,61,62

Table 4 Confusion matrices for *pdfClust* (PC), *Mclust* (MC) and *k-means* (KM) with error rates (ER) for Colon data; in the first column “1” corresponds to tumor samples and “2” to normal.

Real	PC		MC		KM	
	1	2-3	1	2	1	2
1	35	5	29	11	23	17
2	3	19	12	10	6	16
ER:	0.13		0.37		0.37	

our procedure did not recognize any difference between the protocols, cluster 2 and cluster 3 split normal tissues in two groups according to the protocols.

In order to compare our approach to *Mclust*, we carried out a procedure analogous to the one described in Section 2, but using the normal-mixture model both in step (i) and (iii). In the first step, *Mclust* was able to find 369 discriminant genes. We considered the first three principal components of this sub-space for clustering. The procedure found two clusters, with a rather high missclassification error (see Table 4). We also applied a *k-means* algorithm to the entire dataset. The results of the three approaches are shown in Table 4. It can be seen that *k-means*, exploited in the original p -dimensional space, does not perform well. Moreover, *pdfClust* outperforms (in terms of error rate) *Mclust*, if one considers cluster 2 and 3 together as the normal samples.

As stated before, McLachlan et al. [13] studied the same microarray dataset. They selected 446 relevant genes, achieving clusters that seem to recognize the change of protocol in the data structure, but fail to recognize the normal/tumor differences [13]. Nevertheless, they achieved results slightly better than ours (ER= 0.1) considering a particular subspace: they clustered variables in 20 groups and considered only the second group (consisting of 24 genes) to cluster data [13]. Although this approach leads to good results in this example, it seems to be difficult to reproduce the procedure in an unsupervised setting.

Leukaemia data

As stated in [5], the Leukaemia dataset presents two different problems: an easier one, consisting in separating ALL from AML (two-class problem, hereafter) and an harder one, consisting in recognizing also the differences in B-cell and T-cell subclasses (three-class problem).

Again, we considered the strategy described in Section 2. In the variable selection step, *pdfClust* recognizes 313 discriminant genes. Note that the higher number of genes selected with respect to colon data is consistent with the higher difficulty of

Table 5 Confusion matrices for *pdfClust* (PC), *Mclust* (MC) and *k-means* (KM) for Leukaemia data.

Real	PC		MC				KM		
	1	2	1	2	3	4	1	2	3
ALL B-cell	37	1	9	20	9	0	15	0	23
ALL T-cell	5	4	0	0	7	2	7	2	0
AML	4	21	0	2	1	22	1	23	1

the problem. We proceeded by considering the first three principal components of this subspace. *PdfClust* found two clusters, which clearly represent ALL and AML samples, with 4 AML samples classified as ALL and 5 ALL samples classified as AML, leading to a misclassification error rate of 0.125 (Table 5): *pdfClust* is able to solve the two-class problem, but it misses the three-class problem.

In the first step, *Mclust* failed to select relevant genes, recognizing as discriminant among the groups 3,119 out of 3,892 genes. Based on the first three principal components of the subspace spanned by these genes, *Mclust* clustered samples in four groups. We could interpret the merged cluster 1-2 as the ALL B-cell class, and cluster 4 as the AML class, while cluster 3 interpretation is less clear (Table 5). Although *Mclust* is able to find more than two clusters, it fails to distinguish between B-cell and T-cell classes, leading to hardly interpretable clusters.

The leukaemia dataset has been studied by McLachlan et al. as well [13]. The Authors found 2,015 relevant genes after the variable selection step. For the two-class problem, their results were very good (only one sample misallocated), but they failed to solve the three-class problem.

It should be noted that, unlike our algorithm, the procedure used in [13] needs the prior specification of the number of clusters, which is not desirable in an unsupervised learning, especially in cancer tissues classification, where one of the main goal is to find new subclasses of tumors.

6 Conclusions

Model-based approaches to clustering of data have received increasing attention in the last few years, as they provide a sound mathematical-based method. Here, we have discussed a nonparametric density estimation-based algorithm for clustering microarray expression data. Our approach has shown promising results both in simulated data and in two real applications, with surprisingly good computational performances.

In our simulation experiments, we have found that *pdfClust* has performances comparable to those of *Mclust* even in a Gaussian setting (data not shown). Moreover, the gene selection step is much more effective using *pdfClust* than using *Mclust* both in simulated and in real datasets. Here, “effective” means good results in terms of both dimension reduction (e.g. in Leukaemia data *pdfClust* selected 313 genes

versus 3,119 of *Mclust*) and of correct selection (e.g. in GG model the gene selection error rate are 0.08 and 0.77, respectively).

Although we have used here principal components in order to reduce the space dimension, future effort could be done in trying different approaches to this problem. One possibility is to perform a correspondence analysis on the whole dataset and then apply *pdfClust* in the space spanned by the first retained factors. Finally, different clustering techniques could be compared to *pdfClust*, along with *Mclust* and *k-means*, such as the *probabilistic D-clustering* [3].

References

1. Alon, U., Barkai, N., Notterman, D.A., Gish, K., Ybarra, S., Mack, D., Levine, A.J.: Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc. Natl. Acad. Sci. USA* **96**(12), 6745–6750 (1999)
2. Azzalini, A., Torelli, N.: Clustering via nonparametric density estimation. *Stat. Comput.* **17**(1), 71–80 (2007)
3. Ben-Israel, A., Iyigun, C.: Probabilistic D-Clustering. *J. Classif.* **25**(1), 5–26 (2008)
4. Chow, M.L., Moler, E.J., Mian, I.S.: Identifying marker genes in transcription profiling data using a mixture of feature relevance experts. *Physiol. Genom.* **5**(2), 99–111 (2001)
5. Dudoit, S., Fridlyand, J., Speed, T.P.: Comparison of discrimination methods for the classification of tumors using gene expression data. *J. Am. Stat. Assoc.* **97**(457), 77–87 (2002)
6. Fraley, C., Raftery, A.E.: Model-based clustering, discriminant analysis, and density estimation. *J. Am. Stat. Assoc.* **97**(458), 611–631 (2002)
7. Garrett, E.S., Parmigiani, G.: POE: statistical methods for qualitative analysis of gene expression. In: G. Parmigiani, E.S. Garrett, R.A. Irizarry, S.L. Zeger (eds.) *The analysis of gene expression data*, pp. 362–387. Springer (2003)
8. Getz, G., Levine, E., Domany, E.: Coupled two-way clustering analysis of gene microarray data. *Proc. Natl. Acad. Sci. USA* **97**(22), 12,079–12,084 (2000)
9. Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., Coller, H., Loh, M.L., Downing, J.R., Caligiuri, M.A., Bloomfield, C.D., Lander, E.S.: Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science* **286**(5439), 531–537 (1999)
10. Kendzior, C., Newton, M.A., Lan, H., Gould, M.N.: On parametric empirical bayes methods for comparing multiple groups using replicated gene expression profiles. *Stat. Med.* **22**(24), 3899–3914 (2003)
11. Kerr, G., Ruskin, H., Crane, M., Doolan, P.: Techniques for clustering gene expression data. *Comput. Biol. Med.* **38**(3), 283 – 293 (2008)
12. Li, J., Ray, S., Lindsay, B.G.: A nonparametric statistical approach to clustering via mode identification. *J. Mach. Learn. Res.* **8**, 1687–1723 (2007)
13. McLachlan, G.J., Bean, R.W., Peel, D.: A mixture model-based approach to the clustering of microarray expression data. *Bioinformatics* **18**(3), 413–422 (2002)