Physica A 387 (2008) 6310–6318

Contents lists available at ScienceDirect

Physica A

journal homepage: www.elsevier.com/locate/physa

Accuracy and robustness of clustering algorithms for small-size applications in bioinformatics

Pamela Minicozzi [a], Fabio Rapallo [a], Enrico Scalas [a,*], Francesco Dondero [b]

[a] Department of Advanced Sciences and Technology, Università degli Studi del Piemonte Orientale, via Bellini 25g, 15100 Alessandria, Italy

[b] Department of Life and Environmental Science, Università degli Studi del Piemonte Orientale, via Bellini 25g, 15100 Alessandria, Italy

ARTICLE INFO

Article history:
Received 22 May 2008
Received in revised form 23 July 2008
Available online 30 July 2008

PACS:
02.50.Sk
87.18.Wd
87.18.Tt
87.10.Rt

Keywords:
Clustering
DNA microarray
Accuracy
Robustness

ABSTRACT

The performance (accuracy and robustness) of several clustering algorithms is studied for linearly dependent random variables in the presence of noise. It turns out that the error percentage quickly increases when the number of observations is less than the number of variables. This situation is common situation in experiments with DNA microarrays. Moreover, an *a posteriori* criterion to choose between two discordant clustering algorithm is presented.

1. Introduction

Multivariate statistical techniques are an essential tool in many fields of applied science, including Physics, Computer Science, Biology, Medicine, Finance and Economics. In recent years, thanks to the availability of powerful computing tools, such methods have received increasing attention. Among them, *cluster analysis* or *clustering* is used for partitioning available data into groups when prior information is not available or limited. A set of n objects can be allocated into g categories in $\binom{n+g-1}{g-1} = \binom{n+g-1}{n}$ different ways. This number soon becomes very large so that a study by direct enumeration of all possible clusters is no longer tractable. With $n = 20$ objects and $g = 10$ categories, one already has more than ten million possible clusters.

Clustering defines the class of *unsupervised* classification methods [1]. This means that clustering separates a finite data set into a finite number of "natural" categories, where the word *natural* has to be specified according to some measure of closeness between data. Unsupervised classification is opposed to *supervised* classification, where one looks for an accurate characterization of samples generated from some probability distribution with some *a priori* knowledge.

This paper was originally motivated by microarray data analysis. Indeed, Cluster analysis is becoming a major tool in bioinformatics [2–5] and it is widely applied in microarray data analyses: its use is rapidly growing in a wide range of microarray-related problems (see Refs. [6,7]).

[*] Corresponding author. Tel.: +39 0131 360170; fax: +39 0131 360199.
E-mail address: enrico.scalas@mfn.unipmn.it (E. Scalas).

0378-4371/$ – see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.physa.2008.07.026

Table 1
A 6 × 6 matrix from a simulated microarray experiment

	X_1	X_2	X_3	X_4	X_5	X_6
G_1	−0.440	0.563	−0.452	−1.155	1.125	1.162
G_2	−0.531	−0.785	−0.340	−0.793	0.682	1.003
G_3	0.613	1.310	−1.582	−2.209	1.442	1.966
G_4	−0.912	−1.765	−0.491	−0.796	−1.520	−1.820
G_5	1.743	2.185	−1.480	0.003	1.010	1.216
G_6	0.422	0.072	1.604	1.136	−0.064	0.238

The rows $G_1, \ldots, G_6$ denote the genes; the columns $X_1, \ldots, X_6$ denote the tissues.

In a typical microarray experiment, the expression of several thousands of genes is compared in different experimental conditions. The expression is given by the variable

$$x = \log_2\left(\frac{I_R}{I_G}\right),$$

where I_R is the (red) fluorescent intensity coming from reference spots and I_G is the (green) fluorescent intensity coming from treated spots.

After proper normalization and filtering, only a few or at most hundreds of genes result as significantly differentially expressed. This means that x is significantly different from zero. Then, it is useful to apply clustering algorithms in order to detect common patterns of differentially expressed genes. Small groups of genes can be obtained without considering a priori the expression levels, but from a functional analysis through dedicated bioinformatics tools, such as the Gene Ontology annotation [8]. The Gene Ontology is a controlled vocabulary to describe gene and gene product attributes, virtually, in any organism. The Gene Ontology is structured as directed acyclic graphs in which terms are classified in levels and linked through a parent/child relationship. This feature permits to select genes sharing common terms into relative large or small groups depending on the level one is looking at.

From the statistical viewpoint, it is interesting to investigate the behavior of clustering algorithms when the sample size is small. In fact, many statistical procedures dramatically lose accuracy and robustness for small sample sizes.

Moreover, in microarray experiments, data are affected by noise due to measurement errors. Although simple and visually appealing, the performances of clustering algorithms are in general sensitive to noise. Thus, a crucial question is to study their robustness to noise. Some papers in this direction are Refs. [9–11]. The central subject of these works is to perform Monte Carlo simulations to evaluate the behavior of the clustering algorithms. In many cases, the authors define numerical indices to capture the robustness of a clustering algorithm, but such indices are often criticized in subsequent papers.

Having explained our motivations, from now on, we will consider a rather general clustering problem. To illustrate this problem on a synthetic example, let us consider the data in Table 1 with 6 genes and 6 experimental conditions $X_1, \ldots, X_6$.

We applied two different clustering algorithms on the column of the matrix in Table 1. Requiring a final partition with 3 clusters, we have the following results:

- with the single-linkage technique, the final partition is $\{X_1, X_2, X_5, X_6\}$, $\{X_3\}$, $\{X_4\}$;
- with a non-hierarchical technique (PAM), the final partition is $\{X_1, X_2\}$, $\{X_3, X_4\}$, $\{X_5, X_6\}$.

The clustering techniques will be presented in Section 3 with some details. Therefore, a question naturally arises. We have to evaluate what final partition is more likely or, in other words, what algorithm is more accurate in our special case.

In order to answer this question, we can follow two approaches:

- we study the accuracy of the clustering algorithms in several known configurations;
- we make use of the Bayes formula to measure the accuracy of each algorithm.

We will see in the next section that both approaches produce useful information to address this problem.

We study the accuracy and robustness of various clustering algorithms when the distribution of the variables becomes heavy-tailed and for different sample sizes. By accuracy we mean the insensitiveness to noise, while robustness stands for insensitiveness to heavy tails. In particular, we concentrate on small sample sizes. This has been done with a Monte Carlo study where the simple assumption of linearly correlated random variable is used. Moreover, we show how to use the Bayes formula to give an *a posteriori* measure of accuracy for two competing clustering results. We apply this technique to the real data example presented earlier.

The material is organized as follows. In Section 2, some relevant clustering algorithms are described, while in Section 3 we present the design of the simulation study and we discuss the choice of the parameters. The results are summarized in Section 4. In Section 5 we make use of the Bayes rule and of the Monte Carlo algorithms to give a measure of the accuracy when two algorithms produce different partitions and we present a numerical example. Finally, Section 6 is devoted to a discussion of the major findings and of the pointers to future research.