Benchmarking for Clustering Methods Based on Real Data: A Statistical View

Anne-Laure Boulesteix and Myriam Hatz

Abstract In analogy to clinical trials, in a benchmark experiment based on real datasets we can see the considered datasets as playing the role of patients and the compared methods as playing the role of treatments. This view of benchmark experiments, which has already been suggested in the literature, brings to light the importance of statistical concepts such as testing, confidence intervals, power calculation, and sampling procedure for the interpretation of benchmarking results. In this paper we propose an application of these concepts to the special case of benchmark experiments comparing clustering algorithms. We present a simple exemplary benchmarking study comparing two classical clustering algorithms based on 50 high-dimensional gene expression datasets and discuss the interpretation of its results from a critical statistical perspective. The R-codes implementing the analyses presented in this paper are freely available from: http://www.ibe.med.uni-muenchen.de/organisation/mitarbeiter/020_professuren/boulesteix/boulesteixhatz.

1 Introduction

Real data are more complex than simulated data. In practice data never follow well-known distributions. To assess the behavior of data analysis methods in concrete situations of practical relevance, benchmarking on real data is essential.

Whereas supervised learning methods can be evaluated in real data settings based on, e.g., their cross-validation error, there is no obvious criterion to be used to evaluate clustering methods. Quite generally, benchmarking—in particular benchmarking using real data—is a very complex issue in the context of unsupervised learning and to date there still exists no guidance in the literature on how to design and interpret such an experiment. One of the goals of the so-called Task Force on Benchmarking initiated by members of the International Federation of Classification Societies (IFCS) [12] is to provide such guidance.

A.-L. Boulesteix (✉) • M. Hatz
Department of Medical Informatics, Biometry and Epidemiology, University of Munich, München, Germany
e-mail: boulesteix@ibe.med.uni-muenchen.de; myriam.hatz@gmail.com

F. Palumbo et al. (eds.), *Data Science*, Studies in Classification, Data Analysis, and Knowledge Organization, DOI 10.1007/978-3-319-55723-6_6

In the special case where a/the true cluster structure is known, it can be used as a target to be achieved by the clustering method. The agreement between this true cluster structure and the cluster structure output by the method of interest can then be considered as a goodness criterion for evaluating the considered clustering method.

Papers comparing clustering methods typically include simulation studies and an application to a small to moderate number of real datasets. In the present paper, we critically discuss these real data applications from a statistical point of view. In particular, we draw a parallel between benchmark experiments and clinical trials as already suggested by Boulesteix and colleagues [1, 3] for the case of real data and Doove et al. [6] in the context of simulations. In our framework, real datasets play the role of patients, and clustering methods play the role of therapies. With this metaphor in mind, we claim that, in order to make clear statements from real data benchmark experiments, one has to analyze and interpret their results following statistical principles, as illustrated through an exemplary benchmark experiment based on 50 microarray datasets.

Our goal is fourfold: (1) illustrating the variability of benchmarking results across real datasets, (2) propagating statistical thinking in the context of benchmark experiments, where datasets are considered as statistical units, (3) discussing the notion of power in this context, (4) illustrating a possible strategy for the interpretation of benchmark studies based on real datasets through an exemplary study.

The paper is structured as follows. Section 2 briefly presents the clustering methods, data and evaluation criterion used in the exemplary benchmark experiment. The statistical interpretation of the results is given in Sect. 3, including discussions of the concepts of statistical testing, sample size calculation, dependence on datasets' characteristics, and sampling.

2 An Illustrative Benchmark Study: Methods and Data

This section briefly presents the clustering methods, data and evaluation criterion used in the exemplary benchmark experiment.

2.1 Data

The collection of datasets used in our exemplary study was first described by de Souza et al. [5] and used in the context of benchmarking for supervised classification by Boulesteix et al. [4]. It includes 50 clinical gene expression datasets with binary response variable (e.g., diseased vs. healthy), with numbers of patients between $n = 23$ and $n = 286$ and number of variables (genes) between 1098

and 54,680 variables. The datasets can be freely downloaded from the companion website of the paper by Boulesteix et al. [4].

In our study, the interest is in clustering the patients, a task commonly performed in clinical research with the aim, say, to identify typical patient profiles or to discover new disease subtypes. In this context, we would like clustering methods to be able to recover the true cluster structure given by the binary response variable, since it is known to be clinically relevant. Our study includes datasets with binary response variables only to make the comparison of the results across datasets easier.

2.2 Goodness Criterion

As a goodness criterion for clustering methods, we thus simply consider the adjusted Rand index (ARI) [8] measuring the agreement between the true cluster structure (denoted as "partition $\mathscr{C}_{true}$" of $\{1, \ldots, n\}$) defined by the binary response variable and the cluster structure ("partition $\mathscr{C}_M$") output by the clustering method M of interest. The Rand index (RI) can be seen as the proportion of pairs of objects that are either in the same cluster or in different clusters according to both $\mathscr{C}_{true}$ and $\mathscr{C}_M$:

$$RI = \frac{\binom{n_{11}}{2} + \binom{n_{12}}{2} + \binom{n_{21}}{2} + \binom{n_{22}}{2}}{\binom{n}{2}},$$

where $n_{11}, n_{12}, n_{21}, n_{22}$ are the entries of the table showing the numbers of observations from each of the two classes ($Y = 0, 1$) assigned to each of the two clusters ($C = 1, 2$) by clustering method M, see Table 1.

The *adjusted* Rand index (ARI) is an adjusted version of the Rand index accounting for the random agreement and defined as

$$\begin{aligned} ARI &= \frac{RI - mean(RI)}{max(RI) - mean(RI)} \\ &= \frac{\sum_{i,j}\binom{n_{ij}}{2} - \left[\sum_{i=1}^{2}\binom{n_{i.}}{2}\sum_{j=1}^{2}\binom{n_{.j}}{2}\right]/\binom{n}{2}}{\frac{1}{2}\left[\sum_{i=1}^{2}\binom{n_{i.}}{2} + \sum_{j=1}^{2}\binom{n_{.j}}{2}\right] - \left[\sum_{i=1}^{2}\binom{n_{i.}}{2}\sum_{j=1}^{2}\binom{n_{.j}}{2}\right]/\binom{n}{2}}. \end{aligned}$$

Table 1 Partition $\mathscr{C}_M$ output by clustering method M and true clustering $\mathscr{C}_{true}$ given by the binary variable (e.g., diseased vs. healthy)

	$\mathscr{C}_{true}$		
$\mathscr{C}_M$	$Y = 0$	$Y = 1$	Σ
$C = 1$	n_{11}	n_{12}	$n_{1.}$
$C = 2$	n_{21}	n_{22}	$n_{2.}$
Σ	$n_{.1}$	$n_{.2}$	n

2.3 Clustering Methods

In this paper we consider two simple standard clustering methods, since our focus is on issues related to benchmarking and interpretation rather than on the methods themselves. These very widely used methods can be seen as representatives of two important families of clustering methods, namely partitioning methods and hierarchical methods. The first method we consider is partitioning around medoids (PAM) as implemented in the function "pam" of the R package "cluster." The second method is agglomerative hierarchical clustering with euclidean distance as implemented in the function "hclust." These two methods are applied to obtain $K = 2$ clusters, by setting the number of clusters to 2 in "pam" and by cutting the tree in order to obtain two clusters. The choice of $K = 2$ corresponds to the true cluster structure reflected by the binary response variable (note that it would be interesting to also perform analyses with other values of K but this would lead to the problem of the choice of K, which goes beyond the scope of this paper).

3 Statistical Interpretation of Results

In this section, the results of our exemplary benchmark experiment presented in Sect. 2 are discussed from a statistical perspective. Most importantly, we propose to adopt and extend the statistical framework presented by Boulesteix et al. [4] to the context of unsupervised learning.

3.1 Main Results

We obtain the results in the form of a 50×2 matrix containing the *ARI*-values for all 50 datasets and both methods. A straightforward way to visualize the results is to display the *ARI*-values and differences in the form of boxplots as depicted in Fig. 1.

Paired tests can be performed to compare the *ARI*-values of the two methods, as described in Boulesteix et al. [4] in the different case of error rates of classification methods. p-Values of 0.001 and 0.0005 are obtained from the Wilcoxon test and t-test, respectively, whereby the Wilcoxon test seems to be more appropriate considering the skewness of the difference's distribution. In the same vein, one can compute confidence intervals for the median: the bootstrap confidence interval for the median is $(0, 0.053)$ with the percentile method and $(0, 0.046)$ with the bias-corrected accelerated bootstrap method [7].

Beyond statistical tests and the consideration of confidence intervals, further issues related to benchmarking can be advantageously considered from a statistical perspective, in particular in light of clinical trials methodology. They are discussed in the following subsections.

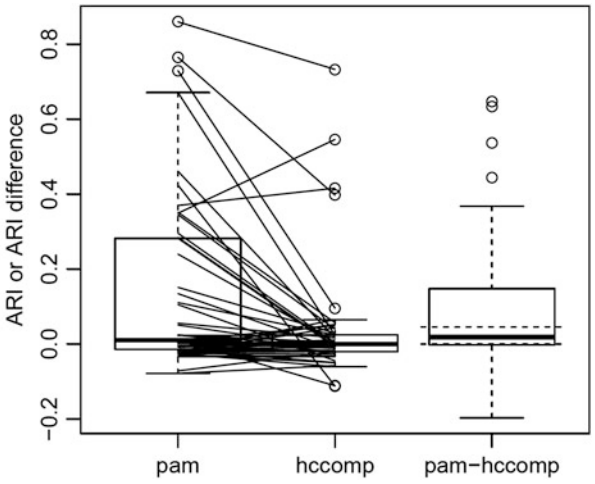

Fig. 1 Adjusted Rand Index (ARI) for the 50 datasets for PAM clustering (*left*) and agglomerative hierarchical clustering with complete linkage (*middle*); difference between these two ARI values (*right*) with the confidence intervals (of the median) obtained by the bias-corrected accelerated bootstrap method represented as *dashed lines*

3.2 Sample Size Calculation

Obviously, the number of datasets included in a benchmark experiment greatly influences the results of the testing procedure. The larger the number of datasets the higher the power to detect differences, and the lower the variance of the estimated difference between the two methods.

To illustrate this issue, we determine the median *ARI*-difference and the p-value of Wilcoxon's test obtained for 1000 random subsets of datasets drawn out of the 50 considered datasets. The corresponding boxplots are displayed in Fig. 2 for different subset sizes ($J = 3$, $J = 5$, $J = 10$, $J = 25$ datasets). As expected, the more datasets one includes in the benchmark experiment, the higher the stability of the median difference in *ARI* and the lower the p-values. If one performs the benchmark experiment based on only $J = 3, 5$, or 10 datasets instead of $J = 50$, the result may look completely different from the results with $J = 50$ datasets. Of note, a number of very large differences (> 0.2) are obtained for $J = 3, 5, 10$. Furthermore, most subsets of size $J = 10$ yield p-values> 0.05.

The notion of power of benchmark experiments in relationship with the number of included datasets can be formally addressed within the statistical testing framework. For simplicity, we assume that the paired t-test is used to compare the two methods. Considering the slightly skewed distribution of the differences between *ARI*-values of the two methods displayed in the right boxplot of Fig. 1, the Wilcoxon is certainly more appropriate. But sample size calculation is essentially an approximative procedure intended to provide orders of magnitude, so considering

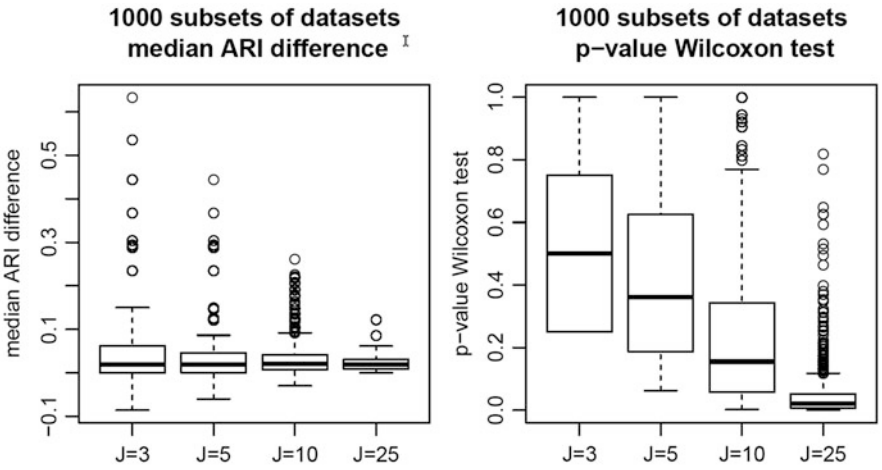

Fig. 2 *Left*: Median difference in *ARI* between the two methods for 1000 subsets of $J = 3, 5, 10, 25$ datasets drawn randomly out of the 50 datasets. *Right*: p-Value of the Wilcoxon test comparing the *ARI*-values of the two methods for 1000 subsets of $J = 3, 5, 10, 25$ datasets drawn randomly out of the 50 datasets

the t-test for the purpose of sample size calculation is acceptable in our context if one keeps in mind that the size has to be slightly increased if the Wilcoxon test is applied instead of the t-test.

The number of observations required to detect a difference of Δ at a significance level of α with a power of $1 - \beta$ using a two-sided paired t-test is approximated by

$$J \approx \frac{(z_{\alpha/2} + z_\beta)^2}{(\Delta/\sigma)^2} \tag{1}$$

where σ denotes the standard deviation of the difference and z_q denotes the q-quantile of the standard normal distribution. Note that this formula is based on the approximation of the Student distribution as standard normal distribution (the exact formula is less easy to apply since it involves the quantiles of the Student distribution, which themselves depend on J).

In our context, σ corresponds to the standard deviation of the difference that is displayed in the right boxplot of Fig. 1 for the 50 datasets. We obtain $\hat{\sigma} = 0.18$. Using Eq. (1), we compute that 25 resp. 102 datasets are required to detect differences of $\Delta = 0.1$ and $\Delta = 0.05$, respectively. Thus, even for a large difference of $\Delta = 0.1$, and for a relatively homogenous set of datasets as considered here (gene expression data, continuous variables, small to moderate sample size), the number of required datasets by far exceeds the size of typical benchmark experiments.

Our results, even if based on a particular example, suggest that it is unrealistic to draw statistically valid conclusions on average superiority of a method over the other based on real datasets without much time and effort. This problem becomes

even more pointed if one does not consider average effects but tries to establish relationships between superiority of methods and datasets' characteristics. This issue is discussed and illustrated in the next section.

3.3 Dependence on Datasets' Characteristics

It can be argued that average superiority over a whole area of application is of poor relevance, since it is expected that the behavior of methods varies a lot depending on datasets' characteristics. Investigating the relationship between datasets' characteristics and methods' performance amounts to examining average superiority within a reduced area defined by particular datasets' characteristics. In this perspective, the issues discussed in the previous sections are also relevant when relating performance/superiority to datasets' characteristics based on real data—and certainly even more since the numbers of datasets are smaller.

It is important to investigate average superiority when elaborating guidelines and establish standard approaches. In an ideal world, methods that establish themselves as standard are those which are superior to other "on average"—even if this is not explicitly tested. Similarly, in an ideal world drugs that are routinely prescribed to patients are those that work best on average according to adequate statistical testing within clinical trials.

However, in the same way as the superior drug may not be the same for two different patients, the superior algorithm may not be the same for different datasets. In both cases, part of these differences might be explained by individual characteristics such as, say, age and sex of the patient and size and number of variables of the dataset, to cite only a few trivial examples. In the same way as a doctor wants to know which drug will best help the patient sitting in front of him, the data analyst wants to know which method performs best for the dataset at hand.

In the clinical context, two strategies have been pursued to address this problem: the search for subgroups in which treatment effects are different, on the one hand, and regression analysis for relating treatment effects to patients' characteristics, on the other hand. In a classical clinical trial with two parallel groups receiving a different treatment, regression analysis is usually performed as follows. The regression model relates the outcome of interest (dependent variable) to the treatment group, the patient's characteristic and their interaction (independent variables). In the context of benchmarking considered here, both methods are applied to all datasets, so the regression model simplifies to a model with the difference of performance as dependent variable and the dataset's characteristic as independent variable.

The search for subgroups can be performed using recursive partitioning methods both in clinical settings [11] and benchmarking settings. This is the approach adopted by Doove et al. [6] in the context of simulation-based benchmarking for clustering methods. In analogy to the term "treatment regime" used in the clinical context, Doove et al. [6] aim at deriving the "optimal data-analytic regime" depending on the dataset's characteristic in the context of benchmark studies.

With real data, however, things are more complex [6], not only because of the lack of straightforward goodness criterion. Firstly, the problem of the limited power is even more of an issue when relating performance to datasets' characteristics than when simply testing average superiority as described in Sect. 3.2. That is because the focus is now essentially on subgroups of datasets. Secondly, datasets' characteristics may be highly correlated, making it different to distinguish their respective effects. In simulation-based benchmarking, some of the relevant datasets' characteristics are controlled by design, hence mitigating this problem. The combination of these two problems makes the investigation of relationships between datasets' characteristics and methods' performance very difficult when using real datasets. On the one hand, the independent effects of the datasets' characteristics can only be assessed by including many of them in the model (keeping in mind that they are not all observable in real data settings!). On the other hand, increasing the number of datasets' characteristics in the model also decreases the power to identify individual effects.

To sum up, claims on the relationships between datasets' characteristics and methods' performance based on real datasets should be formulated very cautiously.

3.4 Sampling Issues and Over-Optimism

In clinical trials precise inclusion criteria for patients are defined before starting patient recruitment, for example, "age>18," "male sex," "no diabetes," etc. All patients fulfilling these criteria are considered for inclusion in the study and asked for their consent. After the data have been collected, it is not allowed to exclude patients from the analysis a posteriori based on their response to therapy.

Such sensible rules should ideally also be adopted in real data-based benchmark studies. Obviously, not all datasets are appropriate to be included in the benchmark study. Or the other way around, a method is not appropriate to all datasets. If some criteria that the dataset has to fulfill to be analyzed with the method are known before performing the benchmark study, candidate datasets should be checked for these criteria and included in the benchmark study only if they fulfill them. All datasets allowed to enter the study should be considered when reporting the results, even those yielding very bad results for the authors' "favorite" method.

Removing these bad datasets from the results has two detrimental consequences: (1) potential important relationships between method performance and datasets' characteristics in the vein of Sect. 3.3 may be overlooked; (2) the overall performance of the "favorite" method may be substantially over-estimated, as outlined theoretically [13] and empirically [9, 10] in the case of supervised learning. By eliminating bad datasets from reporting, one violates rule 4 from the "Ten simple rules to avoid over-optimism in computational research" [2]. This kind of "fishing for datasets" makes the results of real data-based benchmarking even less representative of further datasets.

The definition of inclusion criteria for benchmarking could ideally follow similar principles as in clinical trials. Too strict inclusion criteria lead to study results that are very specific to the considered settings and may not be of broad interest. Conversely, including heterogeneous datasets may make interpretation difficult.

An important difference between benchmarking settings and clinical settings is the "recruitment procedure." For a clinical trial one may, for example, recruit consecutive patients presenting to the hospital with some given symptoms. In the context of benchmarking, however, datasets have to be actively looked for (e.g., in databases or from the companion websites of published papers). This active role of the researcher in the recruitment introduces some arbitrariness and complicates the statistical formalization of the sampling procedure.

There is no straightforward sampling procedure for the population of datasets and the datasets can often not be considered as an *i.i.d.* sample drawn from the population of interest. This may induce biases and dependencies between observations that are difficult to avoid. They should be taken into account when interpreting the results of the benchmarking study. Otherwise, the statistical interpretation of benchmarking may give the readers a false sense of security/scientific correctness. Such issues may be devoted more attention in the context of benchmarking research in the future.

4 Conclusion

Applications to "one or few real datasets" are useful and important. However, they should be considered as illustrative and not representative of what we would obtain with further datasets [1] as long as no statistical inference is performed. Statistical inference requires many datasets and raises important challenges. In particular, there is no straightforward sampling procedure for the population of datasets. Bias is difficult to avoid. In conclusion, results of benchmark experiments based on real datasets should be interpreted with highest caution.

Acknowledgements We thank Sarah Tegenfeldt for language correction and the IFCS Task Force on Benchmarking, in particular to Iven van Mechelen, for very fruitful discussions on the topics of our paper.

References

1. Boulesteix, A.-L.: On representative and illustrative comparisons with real data in bioinformatics: response to the letter to the editor by Smith et al. Bioinformatics **29**(20), 2664–2666 (2013)
2. Boulesteix, A.-L.: Ten simple rules for reducing overoptimistic reporting in methodological computational research. PLOS Comput. Biol. **11**, e1004191 (2015)

3. Boulesteix, A.L., Lauer, S., Eugster, M.J.E.: A plea for neutral comparison studies in computational sciences. PLoS One **8**(4), e61562 (2013)
4. Boulesteix, A.-L., Hable, R., Lauer, S., Eugster, M.J.: A statistical framework for hypothesis testing in real data comparison studies. Am. Stat. **69**, 201–212 (2015)
5. de Souza, B., de Carvalho, A., Soares, C.: A comprehensive comparison of ml algorithms for gene expression data classification. In: Neural Networks (IJCNN), The 2010 International Joint Conference on IEEE, pp. 1–8 (2010)
6. Doove, L., Wilderjans, T., Calcagni, A., van Michelen, I.: Deriving optimal data-analytic regimes from benchmarking studies. Comput. Stat. Data Anal. **107**, 81–91 (2017). http://doi.org/10.1016/j.csda.2016.10.016. http://www.sciencedirect.com/science/article/pii/S0167947316302432
7. Efron, B.: Better bootstrap confidence intervals. J. Am. Stat. Assoc. **82**(397), 171–185 (1987)
8. Hubert, L., Arabie, P.: Comparing partitions. J. Classif. **2**(1), 193–218 (1985)
9. Jelizarow, M., Guillemot, V., Tenenhaus, A., Strimmer, K., Boulesteix, A.-L.: Over-optimism in bioinformatics: an illustration. Bioinformatics **26**(16), 1990–1998 (2010)
10. Macià, N., Bernadó-Mansilla, E., Orriols-Puig, A., Ho, T.K.: Learner excellence biased by data set selection: a case for data characterisation and artificial data sets. Pattern Recogn. **46**(3), 1054–1066 (2013)
11. Seibold, H., Zeileis, A., Hothorn, T.: Model-based recursive partitioning for subgroup analyses. Int. J. Biostat. **12**(1), 45–63 (2016)
12. Steinley, D., van Mechelen, I., IFCS Task Force on Benchmarking, 2015: Benchmarking in cluster analysis: preview of a white paper. Abstract. Conference of the International Federation of Classification Society, Bologna, 6th to 8th July 2015
13. Yousefi, M.R., Hua, J., Sima, C., Dougherty, E.R.: Reporting bias when using real data sets to analyze classification performance. Bioinformatics **26**(1), 68–76 (2010)