Fast and Accurate Estimates for External Clustering Validation Measures

Hugo Sanz-González
Barcelona Supercomputing Center
Barcelona, Spain
hsanzgon@bsc.es

Conrado Martínez
U. Politècnica de Catalunya
Barcelona, Spain
conrado@cs.upc.edu

Reykjavík

Abstract

Many applications need to quantify the similarity between two clusterings of a data set. For example, one good way to assess the quality of a clustering method is to compare its output to a known *ground truth*. applications, it is sometimes worthwhile or even necessary to use estimates in place of exact similarities. Dozens of clustering similarity measures have been invented, yet most of them are based on *pair counting* or on concepts of *information theory*. We investigate how to estimate both kinds of measures accurately and quickly, given a random sample of t data point pairs or of t individual data points.

The resulting estimators either are unbiased or have a bias of $\mathcal{O}(1/t)$, and have a variance of $\mathcal{O}(1/t)$. Our results cover virtually every known pair-counting index, and two of the most popular information-theoretic measures. Regarding computational complexity, the estimates can be obtained in $\mathcal{O}(t)$ time and space. Sometimes these costs include an additional term of order $c \cdot c'$, where c and c' are the respective number of clusters in the two clusterings being compared.

1 Introduction

Let σ be a measure of similarity between two clusterings Π and Π' of a set of n objects. It would be nice if we could take a smaller sample of t of these objects, or of t pairs thereof, and use it to estimate the true similarity value $\sigma(\Pi, \Pi')$. Our goal is to design and analyze statistical estimators $\hat{\sigma}$ of as many clustering similarity measures σ as possible. In particular,

- What is the bias of $\hat{\sigma}(\Pi, \Pi')$? How much does it overestimate or underestimate the true value, on average?
- What is the variance of $\hat{\sigma}(\Pi, \Pi')$, the mean square deviation from its average value?
- How much time and space is needed to compute the estimates?

The answers to these three questions are given in terms of n and t.

2 Applications

External clustering validation. The sheer amount of data in applications like text clustering, customer segmentation, or image processing, makes it difficult, if not impossible, for a person to construct the ground truth clustering of a truly representative set. In such cases, it might be feasible to ask an expert to label only part of the data set, or to answer the question "Should a and b belong to the same cluster?" for a small sample of object pairs $\{a, b\}$. (Using the latter approach, the expert does not even have to determine the individual clusters.) We show how to use this partial information to perform approximate external validation in an efficient and accurate manner.

Consensus clustering. When two or more methods are available for obtaining a "good" clustering of a given data set, it is natural to take into account the similarities between their outputs to reach a *consensus solution*. Most of the similarity measures that appear in the literature can be evaluated in $\mathcal{O}(n + c \cdot c')$ time [1, §4.3]. Even though this cost is generally small compared to that of producing the clusterings in the first place, we may be willing to sacrifice a little bit of accuracy to speed up the consensus clustering process, by using similarity estimates.

3 Results

Proportion indices. Similarity measures like the Rand or Mirkin indices are defined as the fraction p of object pairs that satisfy a certain property. A fairly natural way to estimate p is to take a random sample of t of the $m = \binom{n}{2}$ possible pairs, and to calculate the fraction $\hat{p}$ of them that have the property. Suppose t is a positive random variable; then

$$\mathbb{E}[\hat{p}] = p, \qquad \mathbb{V}[\hat{p}] = p(1-p)\,\mathbb{E}[1/t]$$

when sampling with replacement, and

$$\mathbb{E}[\hat{p}] = p, \qquad \mathbb{V}[\hat{p}] = \frac{mp(1-p)}{m-1}\left(\mathbb{E}\left[\frac{1}{t}\right] - \frac{1}{m}\right)$$

when sampling without replacement. In words, the "straightforward" estimators of a proportion similarity measure are both unbiased and weakly consistent.

Pair-counting measures. A great many clustering similarity measures can be regarded as functions f of a four-dimensional integer vector $\boldsymbol{k} = (k_{11}, k_{10}, k_{01}, k_{00})$, where

$$\begin{aligned}
k_{11} &= \#\text{ of object pairs that fall into the same clusters of } \Pi \text{ and } \Pi',\\
k_{10} &= \#\text{ of object pairs that fall into the same cluster of } \Pi \text{ but not of } \Pi',\\
k_{01} &= \#\text{ of object pairs that fall into the same cluster of } \Pi' \text{ but not of } \Pi,\\
k_{00} &= \#\text{ of object pairs that fall into different clusters of } \Pi \text{ and } \Pi'.
\end{aligned}$$

The adjusted Rand index, the correlation coefficient, and twenty other measures surveyed in [1, Table 6] fit this description.

One attractive way to estimate the similarity $f(\boldsymbol{k})$ between Π and Π' is to select t object pairs at random, to count the numbers $\boldsymbol{X} = (X_{11}, X_{10}, X_{01}, X_{00})$ of pairs in the sample that fall into each category, and to evaluate the statistic $f(\boldsymbol{X})$. If some observation of $\boldsymbol{X}$ makes f diverge or take an indeterminate form, such as $0/0$, we can redefine f to be zero there. The probability that $\boldsymbol{X}$ takes on such problematic values decreases rapidly with increasing t, so we are unlikely to hit them anyway, if t is not too small.

Virtually all pair-counting measures in the literature are positively homogeneous functions of degree 0: they satisfy $f(\lambda\boldsymbol{k}) = f(\boldsymbol{k})$ for all $\lambda > 0$ and all nonzero vectors $\boldsymbol{k}$ of nonnegative real numbers such that $f(\boldsymbol{k})$ exists. We exploit this property, along with the zero-redefinition trick, to prove the following result:

Theorem. *Let $\boldsymbol{X} = (X_{11}, X_{10}, X_{01}, X_{00})$ have the multinomial distribution with positive parameters t and $\boldsymbol{p} = (p_{11}, p_{10}, p_{01}, p_{00})$, or the multivariate hypergeometric distribution with positive parameters t and $\boldsymbol{k} = t\boldsymbol{p}$. Define $q = \sum_{0 \le i,j \le 1} p_{ij}^2$ and $p_\star = \min_{0 \le i,j \le 1} p_{ij}$. Let f be a real positively homogeneous function of degree 0 on an open superset of the hypercube $C = [0\,..\,t]^4$, having continuous second derivatives in the interior of C and in neighborhoods of each point in the range of $\boldsymbol{X}$ where f is nonzero. The bias and variance of $f(\boldsymbol{X})$ as an estimator of $f(\boldsymbol{p})$ are*

$$2a(1-q)/t + \mathcal{O}\left(e^{-tp_\star}\right) \quad \text{and} \quad b(1 + 2\,|f(\boldsymbol{p})|)(1-q)/t + \mathcal{O}\left(e^{-tp_\star}\right)$$

if $\boldsymbol{X}$ is multinomial, and

$$\frac{2c(1-q)(m-t)}{t(m-1)} + \mathcal{O}\left(e^{-tp_\star}\right) \quad \text{and} \quad \frac{d(1 + 2\,|f(\boldsymbol{p})|)(1-q)(m-t)}{t(m-1)} + \mathcal{O}\left(e^{-tp_\star}\right)$$

if $\boldsymbol{X}$ is multivariate hypergeometric. Here $a, b, c, d \ge 0$ are constants not depending on t nor $\boldsymbol{p}$.

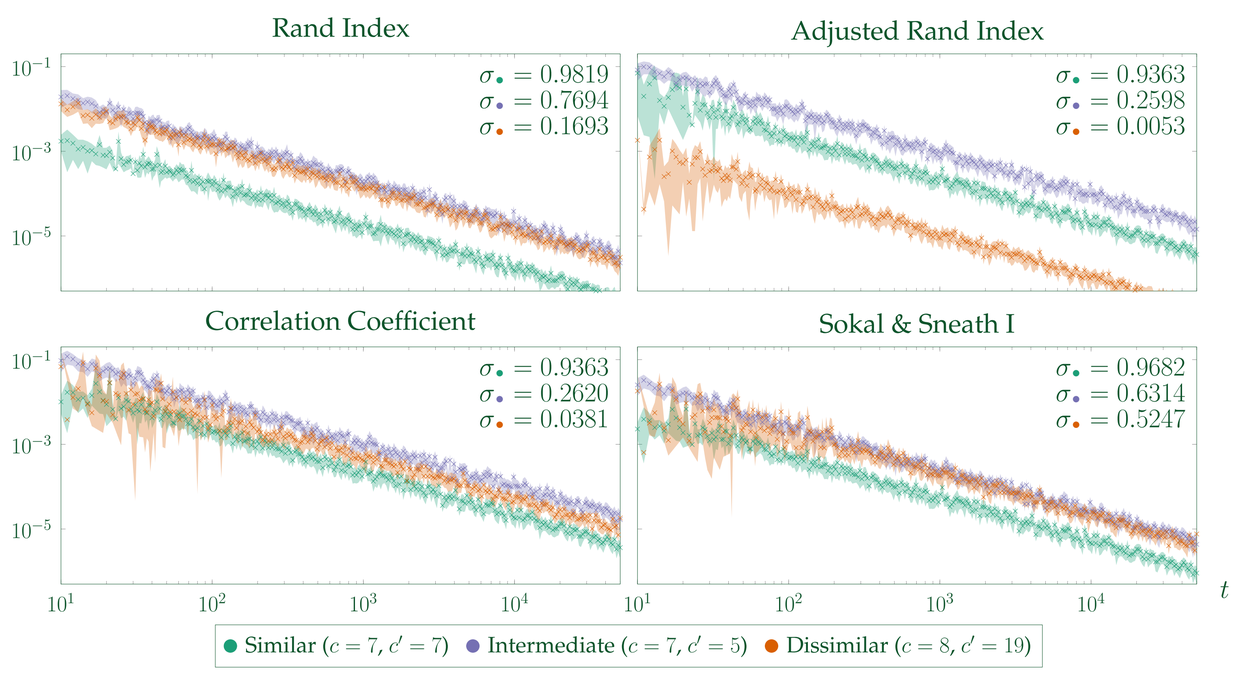
This result applies to nearly all pair-counting similarity measures that are used in applications. It implies that such measures can be estimated with a bias and a variance of $\mathcal{O}(1/t)$, independently of whether the t pairs are sampled with or without replacement. Their estimators are therefore asymptotically unbiased and weakly consistent.

Information-theoretic measures. Another important family of clustering similarity measures can be expressed in terms of the marginal and joint entropies of Y and Y', random variables representing respectively the clusters of Π and Π' that contain a uniformly random object from the partitioned n-set. In particular, the mutual information and the variation of information [2] measures of clustering similarity can be written as sums of these entropies, for which numerous estimators are known. Using the maximum likelihood entropy estimator with the Miller-Madow bias correction [3], for example, yields estimators for these two measures having a bias of $o(1/t)$ and a variance of $\mathcal{O}(1/t)$.

It remains to tackle the normalized and corrected-for-chance information-theoretic measures [4], which cannot be written as linear combinations of entropies. Although they are positively homogeneous of degree 0, they are nondifferentiable at boundary points, so our techniques for pair-counting measures do not apply.

4 Experiments

The following plots show estimates for the mean square error of our pair-counting estimators when given three pairs of clusterings (Π, Π') of varying Rand similarity $\sigma_\bullet$—obtained by a random square dissection process on $n = 10^5$ points—in terms of the number t of object pairs sampled with replacement. Each value is the average of 50 independent trials; the shaded regions show 95% confidence intervals on the mean.



The mean square error of the estimators behaves similarly when the pairs are sampled without replacement. Analogous experimental results for the variation of information and mutual information estimators also agree well with the theory.

5 Conclusions

We have designed statistical estimators for a vast collection of clustering similarity indices. Preliminary experiments show that our bounds on their bias and variance are tight, even in extreme scenarios where the partitions under comparison are very similar or dissimilar. Our results put the estimation of these measures on firm theoretical ground, and help to determine the sample size needed to achieve a target accuracy. In particular, they enable the validation of a clustering algorithm in a semi-supervised manner, where one only needs to annotate a small sample of points or point pairs to get an accurate estimate of $\sigma(\Pi, \Pi_{\text{ground truth}})$.

Our ideas extend beyond measures of clustering similarity, and apply to the similarity between objects of other kinds. We are currently tackling the remaining information-theoretic indices. We would also like to conduct a more thorough experimental validation of our theoretical results, and see how our estimators perform when presented with real-life clusterings.

References

[1] Martijn Gösgens, Alexey Tikhonov, and Liudmila Prokhorenkova. "Systematic Analysis of Cluster Similarity Indices: How to Validate Validation Measures". In: *ICML*. 2021, pp. 3799–3808.

[2] Marina Meilă. "Comparing clusterings—an information based distance". In: *JMVA* 98.5 (2007), pp. 873–895.

[3] Liam Paninski. "Estimation of Entropy and Mutual Information". In: *Neural Comput.* 15.6 (2003), pp. 1191–1253.

[4] Nguyen Xuan Vinh, Julien Epps, and James Bailey. "Information Theoretic Measures for Clusterings Comparison: Variants, Properties, Normalization and Correction for Chance". In: *JMLR* 11.95 (2010), pp. 2837–2854.