Unbiased Similarity Estimators Using Samples

Introduction

Given sets \(A\) and \(B\) from some domain \(U\), random samples \(S_A\) and \(S_B\) from \(A\) and \(B\), respectively, and a similarity measure \(\sigma\),

- Is \((S_A,S_B)\) a good estimator of \(\sigma(A,B)\)?
- If so, is it unbiased? Asymptotically unbiased?
- How does the accuracy of the estimation relate to the size of the samples?
- For which similarity measures do we have good estimators based on random samples?
- Will any random samples do?

The pioneer work of Broder [1] gave some answers for several of these questions in the case of the Jaccard similarity and also for the so-called containment index \(c(A,B)\) that measures how much \(A\) is contained in \(B\). Using random samples of fixed size, one gets unbiased estimators of both measures, the accuracy was not considered.

Applications

Good estimators of similarity using samples will be quite useful in contexts in which we have a large collection of sets \(A_1,A_2,\ldots,A_k\) and we must perform many similarity evaluations \(\sigma(A_i,A_j)\) or \(\sigma(A_i,B)\) in classification tasks or in proximity searches. Substituting complex similarity evaluations by something simpler has been successfully used in many approximate search schemes, like Locality Sensitive Hashing [2,4].

- Extracting a random sample of fixed size \(n\) from \(A\), has cost \(\Theta(|A|\log |A|)
- If \(|A|\) is known then we can set \(k = \log n\), quite intuitively, if \(|A|\) is large we should work with larger samples
- Even if \(|A|\) is not known, one can use a scheme such as Affirmative Sampling [3] which will produce a random sample of (expected) size \(|A|\log |A| / \epsilon < n\), without prior knowledge of \(|A|\), and using extra memory proportional to the size of the sample (not of \(|A|\)).

A Few Technicalities

Many similarity measures are of the form \(\frac{X}{Y}\), where \(X = C \cap B \subseteq B\) that we are going to compute the sets \(A\) and \(B\), \(C\subseteq A\) is the subset of elements of \(C\) that satisfies a certain property. For example,

\[
\text{Jaccard: } \frac{|A \cap B|}{|A \cup B|}
\]

If we are able to construct a random sample \(S_C\) of \(C = S_C \subseteq A \cap B\) and \(S_C \cap C = \text{could be easily computed as well, then } \frac{X}{Y}
\]

\[
\text{we show that it is true asymptotically if variable-size sampling is used, that is, if } S_{|A|} \to n \text{ and } S_{|B|} = \infty \text{ when } |A| \to \infty \text{ and } |B| \to \infty , \text{ by computing all central moments of the estimator.}
\]

Results

Assume we have a hash function \(h : U \to \{0,1\}\), and assume that the probability of collision is negligible (provided that \(h\) has enough bits, we can safely assume that). Given a set \(X\), we denote \(r_X\) the smallest hash value of any element in \(X\); given \(r \in [0,1]\), we denote \(X^{\geq r} = \{ x \in X | x(h(x) \geq r) \}\) the subset of elements in \(X\) with hash value \(\geq r\).

Under reasonable assumptions about the hash function \(h\), \(X^{\geq r}\) is a random sample of \(X\). Any subset of size \(k = |X^{\geq r}|\) from \(X\) is equally likely to be \(X^{\geq r}\). Therefore, to get a random sample of size \(k\) from \(X\), it is enough to collect the \(k\) elements in \(X\) with the largest hash values. There are schemes which allow \(k \to \infty\), where \(n = |X|\) even the need to know \(n\) in advance is unnecessary. For example, Affirmative Sampling, will produce samples of expected size \(\Theta(|A|\log |A|)\) even though \(n\) is not known.

Theorem 1. Let \(\hat{r}\) be any of the similarity measures: Jaccard, Sørensen-Dice, containment coefficient, cosine similarity, Kulczynski 1 (first Kulczynski coefficient), Kulczynski 2 (second) or correlation coefficient.

- Let \(S_A\) and \(S_B\) be random samples of \(A\) and \(B\) such that \(|S_A| = A^{\geq r}\) and \(|S_B| = B^{\geq r}\), and let \(\tau = r^*(S_A,S_B) = \min\{n_A, n_B\} / \min\{|S_A|,|S_B|\}^2\). Then \(\hat{r}(S_A,S_B)\) is an (asymptotically) unbiased estimator of \(\hat{r}(A,B)\), that is,

\[
\hat{r}(S_A,S_B) \sim \hat{r}(A,B)
\]

which implies that \(\tau \to \infty \Rightarrow \text{bias}(\hat{r}(S_A,S_B)) \to 0 \text{ as } |A|, |B| \to \infty .
\]

A similar result holds for other similarity measures like Simpson and Braun-Blanquet.

Conclusions

The similarity of random samples can be used to accurately estimate the similarity of the sets they represent. The samples being of significantly smaller size than the objects, these estimations can be carried out using a tiny fraction of the computational resources one would need to compute the “true” similarity. Some post-processing of the random samples is needed to avoid bias in the estimation, but it does not introduce a serious computational penalty. We have shown that similarity estimation using random samples is possible for many similarity measures between sets, and developed general techniques which might be useful to tackle other new measures not contemplated here. Our careful and solid mathematical analysis (we haven’t just conducted an experimental study) should allow a precise quantitative analysis of the impact of using estimators instead of the “true” similarities in applications.

We are also working on the extension of the ideas and techniques here to other kind of objects like multisets or partitions. For example, we have recently proven that the Rand index (a well known measure of similarity) of two partitions of an \(N\)-element set can be accurately estimated without checking the \((\frac{1}{2})^N\) possible pairs of distinct elements.

References


[3] Braun-Blanquet Cosine Kulczynski 1

Empirical estimates of several similarity measures

The \(x\)-axis in the plots above shows the size of the intersection of two sets \(A = \{t_1, t_2, \ldots, t_m\}\) and \(B = \{s_1, s_2, \ldots, s_m\}\), ranging from 0 (\(r = m = 1\) to \(n = \min(m,n)\) (\(r = 1\)). In the experiments \(m = |A| = 100\) and \(n = |B| = 150\). The red solid lines show the value of \(\hat{r}(A,B)\). The blue dots show the average of \(T = 10\) estimations (sampling 7 times in each set), the blue bars depict the standard variation.

A Coruña, Spain, October 9–11, 2023
