IMPROVING MIN HASH FOR METAGENOMIC TAXONOMIC PROFILING

HOOMAN ZABETI[1], DAVID KOSLICKI[1*]

[1] *Mathematics Department, Oregon State University, Corvallis, OR.*

Abstract. Abstract here.

Keywords: *Min hash, k-mins sketch, metagenomics, taxonomic profiling, taxonomic classification, Jaccard index, containment.*

1. Introduction

(1) Min hash recently has been used to great success on biological data
(2) Mash, Titus' sourmash
(3) originally designed for sets of relatively similar size and appreciable intersection size
(4) metagenomic taxonomic profiling the setup is different: many relatively small database entries, one very large metagenomic sample, very small intersection sizes in general
(5) we modify the min hash paradigm to this particular situation so it can handle a sample of much greater size than the reference database entries.

Min hash is great at comparing sets of similar size. When one set is much larger than the other, the Jaccard index is going to be smaller, which by the Chernoff bounds is where it has a hard time. In metagenomics, the typical paradigm is one very large set (the metagenomic sample) call it B and a bunch of small reference/database sets (call one A). Taking the classical min hash approach means sampling from $A \cup B$. Part a) of Figure 1 demonstrates such a situation while sampling 100 random points of $A \cup B$ and leads to 4 points lying in $A \cap B$. On the other hand, if we sample from just A (instead of $A \cup B$) and have some way to test if a point x is in $A \cap B$, we would get a much better estimate of $|A \cap B|$. Part b) of Figure 1 demonstrates this approach while sampling only 50 points from A and finds 24 points lying in $A \cap B$. They key to our approach is that the membership test $x \in A \cap B$ can be efficiently performed with a bloom filter of B. We analyze the time and space complexity, as well as the accuracy of this approach and find that for parameters typically used in metagenomics, our proposed approach is faster, uses less space, and is more accurate than the classical min hash approach.

2. Methods

Definitions, derivation of mathematical results here.

2.1. Definitions.

(1) Definitions of database entries, query sample, k-mer size, note size disparity
(2) define classic min hash (k-independent version and k-mins version)
(3) define the containment approach

2.2. Min Hash via containment.

Date: April 28, 2017.

[*] Corresponding Author: `david.koslicki@math.oregonstate.edu`.

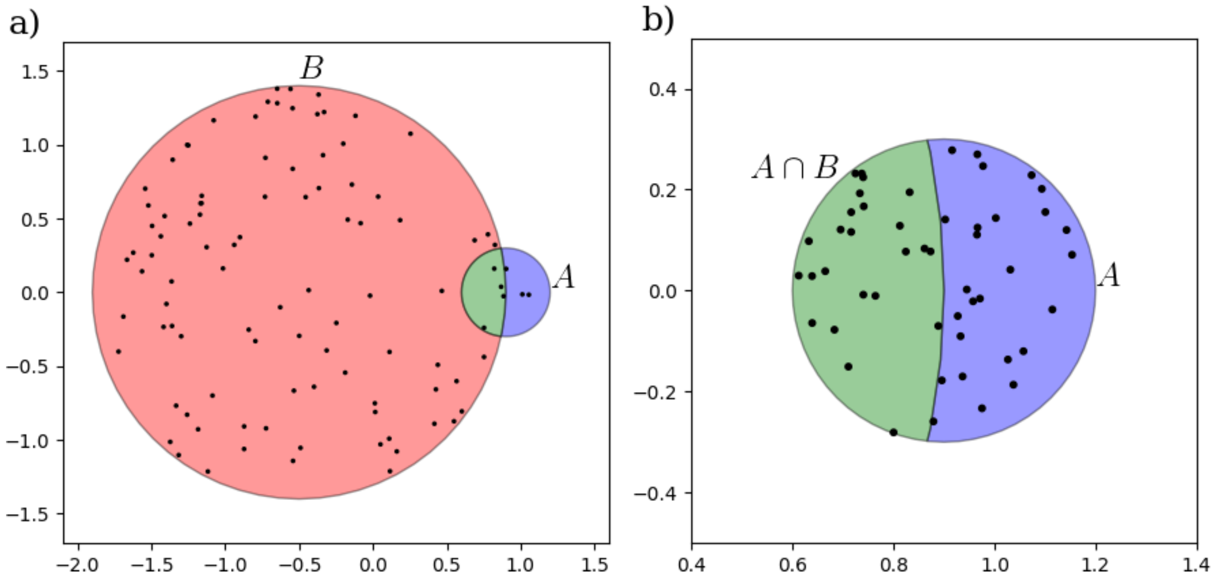

FIGURE 1. Conceptual comparison of classical min hash to the proposed containment approach when estimating the Jaccard index of very different sized sets. a) Sampling 100 points from $A \cup B$ (as is done in the classical min hash approach) leads to finding only 4 elements in $A \cap B$. b) Sampling just 50 points of A and testing if a point $x \in A \cap B$, finds 24 elements in $A \cap B$. This latter approach will be seen to lead to a better estimate of the Jaccard index.

2.3. Time and space complexity.

1. Chernoff bound estimates
2. comparison of number of hashes required for same accuracy
3. time complexity
4. space complexity (all with examples of the numbers in practice).

3. Results

In this section, we compare classic min hash to the proposed method.

3.1. Synthetic data. Here we illustrate the improved accuracy of containment min hash over classical min hash in estimating the Jaccard index. To that end, we generated two random strings w_A and w_B on the alphabet $\{A, C, T, G\}$. We set $|w_A| = 10,000$ and $|w_B| = 15$ to simulate the situation of interest where one wishes to estimate the Jaccard index of two sets of very different size. We then appended a common string w_C of increasing length to each of w_A and w_B so that $\text{Jac}_k(w_A w_C, w_B w_C)$ ranges between 0 and 1. We picked the k-mer size of 11 and utilized a signature size of 100. Figure 2 depicts the comparison of containment min hash with the classical min hash Jaccard estimate on this data and effectively illustrates the results in section 2.2 which proved that the containment approach has a higher probably of being closer to the true Jaccard than the classic approach. The mean and variance of the classic min hash approach on this data was 0.000577 ± 0.001776 while using the containment approach was 0.000717 ± 0.000005 demonstrating a substantial decrease in variance.

3.2. Simulated biological data.

3.3. Real biological data.

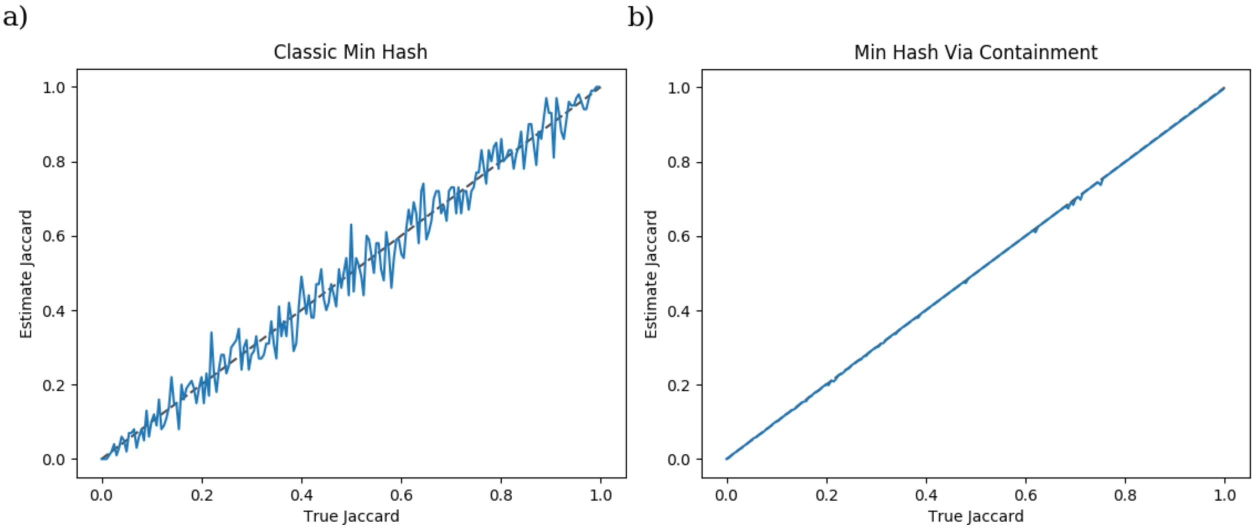

Figure 2. Comparison of containment min hash to the classical min hash estimate of the Jaccard index on synthetic data. Each method utilized the 100 smallest hashes of the murmer3 hash function on the 11-mers of two randomly generated strings with sizes 10,000 and 15 respectively after appending a common substring of increasing size. a) Classical min hash estimate of the Jaccard index. b) The proposed containment min hash on the same data.

4. Discussion