Abstract

We analyze *LSH Forest* [Bawa, Condie, Ganesan 2005]—a popular heuristic for the nearest neighbor search—and show that a careful yet simple modification of it outperforms "vanilla" LSH algorithms. The end result is the first instance of a simple, practical algorithm that provably leverages data-dependent hashing to improve upon data-oblivious LSH. Here is the entire algorithm for the d-dimensional Hamming space. The LSH Forest, for a given dataset, applies a random permutation to all the d coordinates, and builds a *trie* on the resulting strings. In our modification, we further augment this trie: for each node, we store a *constant* number of points close to the mean of the corresponding subset of the dataset, which are compared to any query point reaching that node. The overall data structure is simply several such tries sampled independently. While the new algorithm does not *quantitatively* improve upon the best data-dependent hashing algorithms from [Andoni, Razenshteyn 2015] (which are known to be optimal), it is significantly simpler, being based on a practical heuristic, and is provably better than the best LSH algorithm for the Hamming space [Indyk, Motwani 1998] [Har-Peled, Indyk, Motwani 2012].