Multilinear Hyperplane Hashing

Xianglong Liu† Xinjie Fan† Cheng Deng‡ Zhujin Li† Hao Su§ Dacheng Tao§
†State Key Lab of Software Development Environment, Beihang University, Beijing, China
‡School of Electronic Engineering, Xidian University, Xi’an, China
§Department of Computer Science, Stanford University, Stanford, CA, USA

xlliu@nlsde.buaa.edu.cn chdeng.xd@gmail.com dacheng.tao@uts.edu.au

1. The Proof

Theorem 1 Given a database \(D\) with \(n\) points and a hyperplane query \(\mathcal{P}_w\), if there exists a database point \(x^*\) such that \(d(x^*, \mathcal{P}_w) \leq r\), then with \(\rho = \frac{\ln n}{\ln p}\), using \(n^p\) hash tables with \(\log_{1/p} n\) hash bits, the random multilinear hyperplane hash of an even order is able to return a database point \(x\) such that \(d(x, \mathcal{P}_w) \leq r(1 + \epsilon)\) with probability at least \(1 - \frac{1}{e} - \frac{1}{e} - \epsilon\), \(\epsilon \geq 2\); (2) the query time is sublinear to the entire data number \(n\), with \(n^p \log_{1/p^2} n\) bit generations and \(cn^p\) pairwise distances computation.

This can be completed easily following prior research [1,2].

Proof 1 Denote the number of hash tables to be \(L\). For the \(l\)-th hash table, the proposed MH-Hash algorithm randomly samples \(k\) hash functions \(h_{l,1}^m, \ldots, h_{l,k}^m\) with replacement from \(M_m\), which will generate a \(k\)-bit hash key for each input data vector \(x\). We denote \(x\)'s hash code by \(H_{l,m}^m(x) = [h_{l,1}^m(x), \ldots, h_{l,k}^m(x)]\). The main observation is that using \(L = n^p\) independent hash tables, a \((1 + \epsilon)\)-appropriate nearest neighbor is achieved with a nontrivial constant probability. Moreover, the query (search) time complexity is proved to be sub-linear with respect to the entire data number \(n\).

To complete the proof, we define the following two events \(F_1\) and \(F_2\). It suffices to prove the theorem by showing that both \(F_1\) and \(F_2\) hold with probability larger than 0.5. The two events are defined as follows:

\[ F_1: \text{If there exists a database point } x^* \text{ such that } d(x^*, \mathcal{P}_w) \leq r, \text{ then } H_i^m(x^*) = H_i^m(\mathcal{P}_w) \text{ for some } 1 \leq l \leq L. \]

\[ F_2: \text{Provided with a false alarm set } \mathcal{S} = \{\hat{x} | \hat{x} \in \mathcal{S} \text{ such that } D(\hat{x}, \mathcal{P}_w) > r(1 + \epsilon) \text{ and } \exists l \in [1 : L], H_l^m(\hat{x}) = H_l^m(\mathcal{P}_w)\}, \]

where \(\epsilon > 0\) is the given small constant. Then the set cardinality \(|\mathcal{S}| < cL\).

First, we prove that \(F_1\) holds with probability at least \(1 - \frac{1}{e}\).

Let us consider the converse case that \(H_i^m(x^*) \neq H_i^m(\mathcal{P}_w)\) for all \(l \in [1 : L]\), whose probability is

\[ P[H_i^m(x^*) \neq H_i^m(\mathcal{P}_w), \forall l \in [1 : L]] = (1 - P[H_i^m(x^*) \neq H_i^m(\mathcal{P}_w)])^L \]

\[ = (1 - (1 - p_1 k)^L = (1 - \frac{1}{e} n^{\rho})^L = (1 - n^{-\rho} n^{\rho})^L \]

\[ = ((1 - n^{-\rho}) - n^{-\rho})^{-1} \leq \frac{1}{e}, \]

where inequality (1) follows the inequality \((1 - n^{-\rho}) - n^{-\rho} \geq e\). Herewith we derive

\[ P[H_i^m(x^*) = H_i^m(\mathcal{P}_w), \exists l \in [1 : L]] = 1 - P[H_i^m(x^*) \neq H_i^m(\mathcal{P}_w), \forall l \in [1 : H]] \]

\[ \geq 1 - \frac{1}{e}. \]

Second, we prove that \(F_2\) holds with probability at least \(1 - \frac{1}{e}\).

For every false alarm point \(\hat{x}\) conforming to \(D(\hat{x}, \mathcal{P}_w) > r(1 + \epsilon)\), in any hash table \(l \in [1 : L]\) we have

\[ P[H_l^m(\hat{x}) = H_l^m(\mathcal{P}_w)] < p_2 k = (p_2)^{\log_{1/p^2} n} = \frac{1}{n}. \]

Therefore the expected number of false alarm points, which fall into the same hash bucket with the query \(\mathcal{P}_w\) in hash table \(l\), is smaller than \(n \times 1/n = 1\). Immediately, we
conclude $E[|S|] < L$. Subsequently, we further apply Markov's inequality to derive the following result:

$$P[|S| \geq cL] \leq \frac{E[|S|]}{cL} < \frac{L}{cL} = \frac{1}{c},$$

which leads to

$$P[|S| < cL] = 1 - P[|S| \geq cL] > 1 - \frac{1}{c}.$$

Third, we prove that $F_1$ and $F_2$ simultaneously hold with probability at least $1 - \frac{1}{c} - \frac{1}{e}$.

References
