Hash Bit Selection: a Unified Solution for Selection Problems in Hashing

Xianglong Liu1, Junfeng He2,3, Bo Lang1, and Shih-Fu Chang2
1Beihang University, 2Columbia University, 3Facebook

1. Overview

- **Problem:** to apply hashing techniques successfully, there are several important issues remaining open in selecting features, hashing algorithms, parameter settings, kernels, etc.
- **Motivation:** similar to feature selection, give a unified solution that can directly select the most desirable subset of hash bits from different bit sources, targeting the specific scenario.
- **Scenarios:** hashing with multiple features, multiple hashing algorithms, multi-bit hashing algorithms, etc.

2. Hash Bit Selection

- **Goal:** to exploit a small bit subset of size \( l \) from the pooled \( L \) types of bits with index set \( V = \{1, \ldots, L\} \)
- **Criteria:** two properties critical for compact hash codes
  - A. similarity preservation: \( \eta_i = \exp(-\gamma y_i^T G y_i) \)
  - B. mutual independence: \( \alpha_{ij} = \exp\left( -\sum_{h,v} p(h,v) \log \frac{p(h,v)}{p(h)p(v)} \right) \)
- **Formulation:** a quadratic programming (QP) solved efficiently

\[
\max_{\mathbf{x} \in \Omega} \mathbf{x}^T A \mathbf{x} \\
\text{s.t.} \quad \mathbf{x} \in \Omega \\
\Omega = \{ \mathbf{x} \in \{0,1\}^L : |\mathbf{x}|_0 = l \} \\
\mathbf{x} \in \mathbb{R}^L : \mathbf{x} \geq 0 \text{ and } 1^T \mathbf{x} = 1 \}
\]

Cohesivity between bits: \( A = I_{LxL} \), satisfying:

\( \tilde{A}_{ij} \geq 0, \tilde{A}_{ij} = A_{ij}, \text{ and } \tilde{A}_{ij} \propto \pi_i, \pi_j, \text{ and } a_{ij} \)

3. Theoretic Analysis

- **Graph representation:** represent the pooled bits as a vertex-weighted and undirected edge-weighted graph \( G = (V, E, A, \pi) \)
- **Normalized dominant set (NDomSet):** bit selection is equivalent to the discovery of a normalized dominant set that has high internal vertex weights and edge weights
- **Induced vertex weights:** naturally score each vertex, and a NDomSet has \( \omega_S(i) > 0 \)

4. Experimental Results

- **with multiple features:** CIAFR-10 and NUS-WIDE
- **Over multiple hashing algorithms:** GIST-1M
- **Over multi-bit hashing algorithms:** GIST-1M

Our proposed method performs significantly better than existing approaches

5. Summary and Conclusions

- Propose a generic bit selection unifying various scenarios
- Consider two important criteria tailored for hashing performance
- Formulate the problem as a quadratic programming and reveal its nature via the normalized dominant set
- More information http://www.nlsde.buaa.edu.cn/~xlliu