Geometric Constraints CIS 6930 Fall 2006

Lecture :LSH Based on p-Stable Distribution

Lecturer: Dr. Meera Sitharam Scribe: Heping Gao

The problem will be discussed here is: given objects represented as points in $\mathcal{R}^d$ and a distance metric which is used to measure similarity of objects, how to perform indexing or similarity searching for query objects. The dimensionality d ranges anywhere from tens to thousands. Note that the low-dimensional case (say the dimensionality d equal to 2 or 3) is well-solved, so the main issues is that of dealing with a large number of dimensions.

The key idea of *locality-sensitive hash (LSH)* is to hash the points using several hash functions so as to ensure that, for each function, the probability of collision is much higher for objects which are close to each other than for those which are far apart. Then, one can determine near neighbors by hashing the query point and retrieving elements stored in buckets containing that point.

Since the LSH is a hashing-based scheme, it can be naturally extended to the *dynamic* setting, i.e., when insertion and deletion operations also need to be supported. This avoids the complexity of dealing with tree structures when the data is dynamic.

This paper presented a novel version of the LSH algorithm. It works for the $(R,c)-NN$ problem, where the goal is to report a point within distance R from q. The advantages of the new algorithm is:

- For the l_2 norm, its query time is $O(dn^{\rho(c)}\log n)$, where $\rho(c) < 1/c$ for $c \in (1, 10]$.
- It is simple and quite easy to implement.
- It works for any l_p norm, as long as $p \in (0, 2]$. Specifically, it is shown in the paper that for any $p \in (0, 2]$ and $\gamma > 0$ there exists an algorithm for $(R,c)-NN$ under l_d^p which uses $O(dn+n^{1+\rho})$ space, with query time $O(n^{\rho}\log_{1/\gamma} n)$, where $\rho \leq (1+\gamma)\cdot max(\frac{1}{c^p}, \frac{1}{c})$.

Let $M = (X, d)$ be any metric space, and $v \in X$. The ball of radius r centered at v is defined as $B(v, r) = \{q \in X\}|d(v, q) \leq r$.

For a domain S of the points set with distance measure D, an LSH family is defined as:

Definition 1 *A family $\mathcal{H} = \{h : S \to U\}$ is called $(r_1, r_2, p_1, p_2) - sensitive$ for D if for any $v, q \in S$*

- *if $v \in B(q, r_1)$ then $Pr_{\mathcal{H}}[h(q) = h(v)] \geq p_1$.*
- *if $v \notin B(q, r_2)$ then $Pr_{\mathcal{H}}[h(q) = h(v)] \leq p_2$.*

In order for a locality-sensitive hash (LSH) family to be useful, it has to satisfy inequalities $p_1 > p_2$ and $r_1 < r_2$.

Theorem 2 *Suppose there is a $(R, cR, p_1, p_2) - sensitive$ family $\mathcal{H}$ for a distance measure D. Then there exists an algorithm for $(R, c) - NN$ under measure D which use $O(dn + n^{1+\rho})$ space, with query time dominated by $O(n^{\rho})$ distance computations, and $O(n^p \log_{1/p_2} n)$ evaluations of hash functions from $\mathcal{H}$, where $\rho = \frac{\ln 1/p_1}{\ln 1/p_2}$.*

The following notes will show how to get the hash family.

Definition 3 *A distribution D over $\mathcal{R}$ is called p-stable, if there exists $p \geq 0$ such that for any n real numbers $v_1 \cdots v_n$ and i.i.d. variables $X_1 \cdots X_n$ with distribution D, the random variable $\sum_i v_i X_i$ has the same distribution as the variable $(\sum_i |v_i^p|)^{1/p} X$, where X is a random variable with distribution D.*

Each hash function is given as $h_{a,b}(v) = \lfloor \frac{a \cdot v + b}{r} \rfloor$ where a is a d-dimensional vector with entries chosen independently from a p-stable distribution and b is a real number chosen uniformly from the range $[0, r]$. So each hash function is indexed by random a and b. Since stable distribution exist for any $p \in (0, 2]$, we can find hash family for all $p \in (0, 2]$ by this way.

Now let's have a look at the property of hash function $h_{a,b}(v)$. For two vectors v_1, v_2, let $c = \|v_1 - v_2\|_p$. For a random vector a whose entries are drawn from a p-stable distribution, $a \cdot v_1 - a \cdot v_2$ is distributed as cX where X is a random variable drawn from a p-stable distribution. Since b is drawn uniformly from $[0, r]$ it is easy to see that

$p(c) = Pr_{a,b}[h_{a,b}(v_1) = h_{a,b}(v_2)] = \int_0^r \frac{1}{c} f_p(\frac{t}{c})(1 - \frac{t}{r})dt$

Here $f_p(t)$ denotes the probability density function of the absolute value of the p-stable distribution.

As per Definition 1, the family of hash functions above is $(r1, r1 * c, p(1), p(c)) -$ *sensitive*.

References

[1] Datar, M., and Immorlica, N. and Indyk, P. and Mirrokni, V. "Locality-Sensitive hashing Scheme Based on p-Stable Distribution," *www.mit.edu/~mirrokni/pstable.ps.*