Correlated locality-sensitive hashing*

Rasmus Pagh

IT University of Copenhagen, Denmark

Abstract. After an introduction to the area, we consider a new construction of locality-sensitive hash functions for Hamming space that is *covering* in the sense that is it guaranteed to produce a collision for every pair of vectors within a given radius r. The construction is *efficient* in the sense that the expected number of hash collisions between vectors at distance cr, for a given $c > 1$, comes close to that of the best possible data independent LSH without the covering guarantee, namely, the seminal LSH construction of Indyk and Motwani (FOCS '98). The efficiency of the new construction essentially *matches* their bound if $\log(n)/(cr)$ is integer, where n is the number of points in the data set, and differs from it by at most a factor $\ln(4) < 1.4$ in the exponent for larger values of cr. As a consequence, LSH-based similarity search in Hamming space can avoid the problem of false negatives at little or no cost in efficiency.

* The research leading to these results has received funding from the European Research Council under the European Union's Seventh Framework Programme (FP7/2007-2013) / ERC grant agreement no. 614331.