Beating brute-force: Improved algorithms for finding correlations, and related problems.

Gregory Valiant

Microsoft Research

ABSTRACT

Perhaps the most basic type of structure in a dataset is correlation. Computationally, how quickly can correlated variables be found? One can certainly brute-force search through all pairs of variables, and for each pair, the correlation can be approximated very efficiently. But is there a *sub-quadratic* time algorithm for finding pairs of correlated variables?

More generally, suppose one has a dataset where each data example has a label, given as some function of a small number of the variables. If we have n total variables, perhaps there is a small number, $k = 2, 3, 4, \ldots$, of *relevant* variables which can be used to predict the labels. Such a function is termed a *k-junta*. How quickly can one find this set of relevant variables? As above, one could perform a brute-force search over all possible subsets of size k, taking time roughly $O(n^k)$. Can one find the set of relevant variables significantly more efficiently?

We give affirmative answers to both questions. We show that a pair of ρ-correlated variable can be found in a set of n otherwise random Boolean variables in time $O\left(n^{1.6} poly(1/\rho)\right)$. This improves upon the $O(n^{2-O(\rho)})$ runtime given by *locality sensitive hashing* and related approaches [1]. Applications and extensions of our basic approach yield algorithms with improved asymptotic runtimes for several other problems, including learning k-juntas, learning sparse parity with noise, and computing the approximate closest pair of points, in both Euclidean and Boolean settings.

This work will appear at FOCS'12 as [2].

BODY

One can find correlated variables in time $O(n^{1.6})$. Similar ideas give faster algorithms for Closest-Pair, and learning juntas/parities.

REFERENCES

[1] P. Indyk and R. Motwani. Approximate nearest neighbors: towards removing the curse of dimensionality. In *Proceedings of the ACM Symposium on Theory of Computing (STOC)*, 1998.

[2] G. Valiant. Finding correlations in subquadratic time, with applications to learning parities and juntas. In *IEEE Symposium on Foundations of Computer Science (FOCS) (to appear)*, 2012.

Volume 1 of Tiny Transactions on Computer Science