Hashing Algorithms for Large-Scale Search and Learning

Cun-Hui Zhang

Rutgers University

4:30 pm, 6/27/2018, 1195 Bordeaux Dr, Sunnyvale, CA 94089

ABSTRACT

The talk begins with an introduction of minwise hashing (minhash), which has been widely used in the search industry. The variant named "b-bit minwise hashing" can be seamlessly integrated into machine learning algorithms for training ultra-large and ultra-high dimensional models with binary (0/1) data. The bottleneck of the original b-bit minhash lies in the expensive preprocessing cost. In NIPS 2012, our work on "one permutation hashing" successfully and rigorously reduced the preprocessing cost by a factor of 100 or more. In fact, in many practical applications such as CTR (click-through rate) models, with b-bit one permutation hashing, no additional preprocessing is needed and the feature signature size could be substantially reduced. The talk then moves on to the more recent work (WWW 2017) on the theoretical property of "generalized min-max similarity" for non-binary data and the related hashing method named "generalized consistent weighted sampling". These methods are promising for producing simple machine learning models with accuracies comparable to (or in some cases better than) boosted trees and deep nets.

Bio: Cun-Hui Zhang is Distinguished Professor of Statistics at Rutgers University. His research interests include machine learning, high-dimensional data, bootstrap, data sketch and hashing, nonparametric methods, multivariate analysis, survival data, functional MRI, closed loop diabetes control, and network tomography. He is a Fellow of the Institute of Mathematical Statistics and a Fellow of American Statistical Association. He is Editor of Statistical Science and serves on the editorial boards of Annals of Statistics, Bernoulli, Statistica Sinica, and Statistics Surveys.