Computer Science > Learning

b-Bit Minwise Hashing for Large-Scale Linear SVM

Ping Li, Joshua Moore, Christian Konig

(Submitted on 23 May 2011)

In this paper, we propose to (seamlessly) integrate b-bit minwise hashing with linear SVM to substantially improve the training (and testing) efficiency using much smaller memory, with essentially no loss of accuracy. Theoretically, we prove that the resemblance matrix, the minwise hashing matrix, and the b-bit minwise hashing matrix are all positive definite matrices (kernels). Interestingly, our proof for the positive definiteness of the b-bit minwise hashing kernel naturally suggests a simple strategy to integrate b-bit hashing with linear SVM. Our technique is particularly useful when the data can not fit in memory, which is an increasingly critical issue in large-scale machine learning. Our preliminary experimental results on a publicly available webspam dataset (350K samples and 16 million dimensions) verified the effectiveness of our algorithm. For example, the training time was reduced to merely a few seconds. In addition, our technique can be easily extended to many other linear and nonlinear machine learning applications such as logistic regression.

Subjects: **Learning (cs.LG)**; Applications (stat.AP); Computation (stat.CO); Machine Learning (stat.ML)

Cite as: **arXiv:1105.4385 [cs.LG]**

(or **arXiv:1105.4385v1 [cs.LG]** for this version)

Submission history

From: Ping Li [view email]

[v1] Mon, 23 May 2011 01:56:24 GMT (81kb,S)

Which authors of this paper are endorsers?