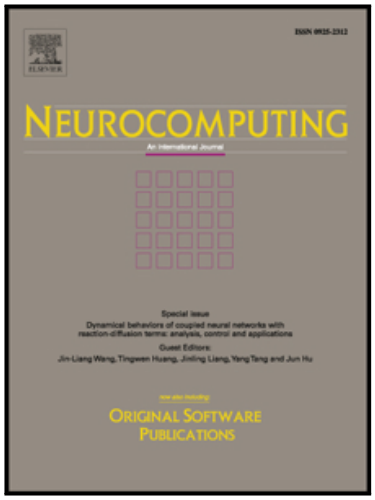
Communicated by Dr. Sato-Ilic Mika

Accepted Manuscript



Large-scale k-means clustering via variance reduction

Yawei Zhao, Yuewei Ming, Xinwang Liu, En Zhu, Kaikai Zhao, Jianping Yin

PII: S0925-2312(18)30458-2
DOI: 10.1016/j.neucom.2018.03.059
Reference: NEUCOM 19491

To appear in: *Neurocomputing*

Received date: 8 November 2017
Revised date: 20 January 2018
Accepted date: 29 March 2018

Please cite this article as: Yawei Zhao, Yuewei Ming, Xinwang Liu, En Zhu, Kaikai Zhao, Jianping Yin, Large-scale k-means clustering via variance reduction, *Neurocomputing* (2018), doi: 10.1016/j.neucom.2018.03.059

Large-scale k-means clustering via variance reduction

Yawei Zhao, Yuewei Ming, Xinwang Liu, En Zhu

National University of Defense Technology, Changsha, Hunan, China

Kaikai Zhao

Naval Aeronautical University, Yantai, Shandong, China

Jianping Yin

Dongguan University of Technology, Dongguan, Guangdong, China

Abstract

With the increase of the volume of data such as images in web, it is challenging to perform k-means clustering on millions or even billions of images efficiently. One of the reasons is that k-means requires to use a batch of training data to update cluster centers at every iteration, which is time-consuming. Conventionally, k-means is accelerated by using one or a mini-batch of instances to update the centers, which leads to a bad performance due to the stochastic noise. In the paper, we decrease such stochastic noise, and accelerate k-means by using variance reduction technique. Specifically, we propose a position correction mechanism to correct the drift of the cluster centers, and propose a variance reduced k-means named VRKM. Furthermore, we optimize VRKM by reducing its computational cost, and propose a new variant of the variance reduced k-means named VRKM++. Comparing with VRKM, VRKM++ does not have to compute the batch gradient, and is more efficient. Extensive empirical studies show that our methods VRKM and VRKM++ outperform the state-of-the-art method, and obtain about 2× and 4× speedups for large-scale clustering, respectively. The source code is available at https://github.com/YaweiZhao/VRKM_sofia-ml.