

Accepted Manuscript

Large-scale k-means clustering via variance reduction

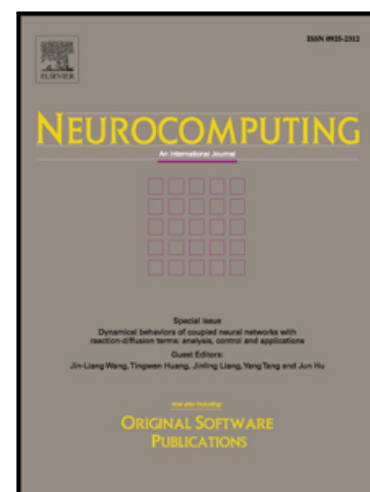
Yawei Zhao, Yuewei Ming, Xinwang Liu, En Zhu, Kaikai Zhao,
Jianping Yin

PII: S0925-2312(18)30458-2
DOI: [10.1016/j.neucom.2018.03.059](https://doi.org/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](https://doi.org/10.1016/j.neucom.2018.03.059)



This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

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 $2\times$ and $4\times$ speedups for large-scale clustering, respectively. The source code is available at https://github.com/YaweiZhao/VRKM_sofia-ml.

Download English Version:

<https://daneshyari.com/en/article/6863752>

Download Persian Version:

<https://daneshyari.com/article/6863752>

[Daneshyari.com](https://daneshyari.com)