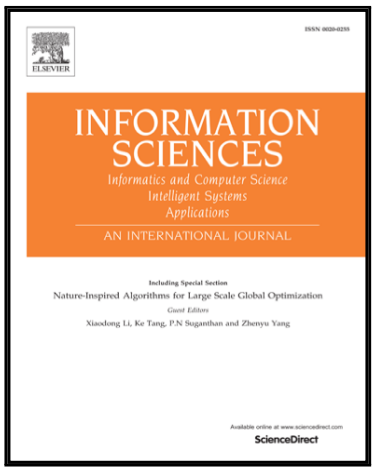


# Voting-based Instance Selection from Large Data Sets with MapReduce and Random Weight Networks

Junhai Zhai, Xizhao Wang, Xiaohe Pang

PII: S0020-0255(16)30508-4
DOI: 10.1016/j.ins.2016.07.026
Reference: INS 12352

To appear in: *Information Sciences*

Received date: 16 May 2015
Revised date: 9 June 2016
Accepted date: 6 July 2016

Please cite this article as: Junhai Zhai, Xizhao Wang, Xiaohe Pang, Voting-based Instance Selection from Large Data Sets with MapReduce and Random Weight Networks, *Information Sciences* (2016), doi: 10.1016/j.ins.2016.07.026

# Voting-based Instance Selection from Large Data Sets with MapReduce and Random Weight Networks

Junhai Zhai[a,b,*], Xizhao Wang[c], Xiaohe Pang[d]

[a] *Key Lab. of Machine Learning and Computational Intelligence, College of Mathematics and Information Science, Hebei University, Baoding, 071002, Hebei, China*
[b] *College of Mathematics, Physics and Information Engineering, Zhejiang Normal University, Jinhua 321004, China*
[c] *College of Computer Science and Software, Shenzhen University, Shenzhen 518060, China*
[d] *College of Computer Science and Technology, Hebei University, Baoding, 071002, Hebei, China*

**Abstract**

Instance selection is an important preprocessing step in machine learning. By choosing a subset of a data set, it achieves the same performance of a machine learning algorithm as if the whole data set is used, and it enables a machine learning algorithm to be feasible for and to work effectively with large data sets. Based on voting mechanism, this paper proposes a large data sets instance selection algorithm with MapReduce and random weight networks (RWNs). Firstly, the proposed algorithm employs the Map of MapRecuce to partition the large data sets into some small subsets, and deploys them to different cloud computing nodes. Secondly, the informative instances are selected in parallel with an instance selection algorithm. Thirdly, the Reduce of MapRecuce is used to collect the selected instances from different cloud computing nodes and a selected instance subset is obtained. The previous three processes are repeated $p$ times ($p$ is a parameter defined by the user), and $p$ instance subsets are obtained. Finally, the voting method is used to select the most informative instances from the $p$ subsets. The random weight network classifier is trained with the selected instance subset, and the testing accuracy is verified on the testing set. The proposed algorithm is experimentally compared with three state-of-the-art approaches which are CNN, ENN and RNN. The experimental results show that the proposed

*E-mail: mczjh@126.com