Mach Learn (2013) 92:195–220
DOI 10.1007/s10994-013-5373-4

Massively parallel feature selection: an approach based on variance preservation

Zheng Zhao · Ruiwen Zhang · James Cox · David Duling · Warren Sarle

Received: 17 November 2012 / Accepted: 27 April 2013 / Published online: 22 May 2013

Abstract Advances in computer technologies have enabled corporations to accumulate data at an unprecedented speed. Large-scale business data might contain billions of observations and thousands of features, which easily brings their scale to the level of terabytes. Most traditional feature selection algorithms are designed and implemented for a centralized computing architecture. Their usability significantly deteriorates when data size exceeds tens of gigabytes. High-performance distributed computing frameworks and protocols, such as the Message Passing Interface (MPI) and MapReduce, have been proposed to facilitate software development on grid infrastructures, enabling analysts to process large-scale problems efficiently. This paper presents a novel large-scale feature selection algorithm that is based on variance analysis. The algorithm selects features by evaluating their abilities to explain data variance. It supports both supervised and unsupervised feature selection and can be readily implemented in most distributed computing environments. The algorithm was implemented as a SAS High-Performance Analytics procedure, which can read data in distributed form and perform parallel feature selection in both symmetric multiprocessing mode (SMP) and massively parallel processing mode (MPP). Experimental results demonstrated the superior performance of the proposed method for large scale feature selection.

Keywords Feature selection · Model selection · Parallel processing · Big-data

Editors: Tijl De Bie and Peter Flach

Z. Zhao (✉) · R. Zhang · J. Cox · D. Duling · W. Sarle
SAS Institute Inc., 600 Research Drive, Cary, NC 27513, USA
e-mail: zheng.zhao@sas.com

R. Zhang
e-mail: ruiwen.zhang@sas.com

J. Cox
e-mail: james.cox@sas.com

D. Duling
e-mail: david.duling@sas.com

W. Sarle
e-mail: warren.sarle@sas.com