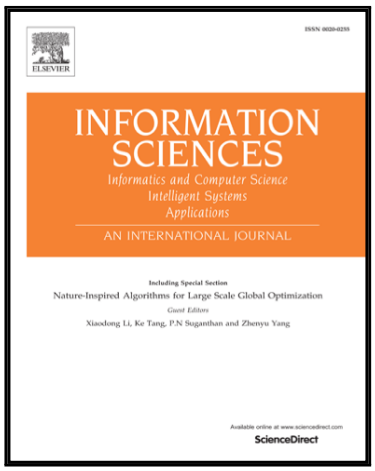

A Pearson's correlation coefficient based decision tree and its parallel implementation

Yashuang Mu, Xiaodong Liu, Lidong Wang

PII: S0020-0255(17)31176-3
DOI: 10.1016/j.ins.2017.12.059
Reference: INS 13353

To appear in: *Information Sciences*

Received date: 11 March 2017
Revised date: 28 December 2017
Accepted date: 29 December 2017

Please cite this article as: Yashuang Mu, Xiaodong Liu, Lidong Wang, A Pearson's correlation coefficient based decision tree and its parallel implementation, *Information Sciences* (2017), doi: 10.1016/j.ins.2017.12.059

A Pearson's correlation coefficient based decision tree and its parallel implementation

Yashuang Mu[a], Xiaodong Liu[a,*], Lidong Wang[b]

[a] *School of Control Science and Engineering, Faculty of Electronic Information and Electrical Engineering, Dalian University of Technology, Dalian 116024, P. R. China.*
[b] *Department of Mathematics, Dalian Maritime University, Dalian 116026, P. R. China.*

Abstract

In this paper, a Pearson's correlation coefficient based decision tree (PCC-Tree) is established and its parallel implementation is developed in the framework of Map-Reduce (MR-PCC-Tree). The proposed methods employ Pearson's correlation coefficient as a new measure of feature quality to confirm the optimal splitting attributes and splitting points in the growth of decision trees. Besides, the proposed MR-PCC-Tree adopts Map-Reduce technology to every component during the decision trees learning process for parallel computing, which mainly consists of a parallel Pearson's correlation coefficient based splitting rule and a parallel splitting data method. The experimental analysis is conducted on a series of UCI benchmark data sets with different scales. In contrast to several traditional decision tree classifiers including BFT, C4.5, LAD, SC and NBT on 17 data sets, the proposed PCC-Tree is no worse than the traditional models as a whole. Furthermore, the experimental results on other 8 data sets show the feasibility of the proposed MR-PCC-Tree and its good parallel performance on reducing computational time for large-scale data classification problems.

Keywords: Decision trees, Parallel computing, Pearson's correlation coefficient, Map-Reduce

*Corresponding author
Email addresses: muyashuang324@126.com (Yashuang Mu), xdliuros@dlut.edu.cn (Xiaodong Liu), ldwang@hotmail.com (Lidong Wang)