FCNN-MR: A Parallel Instance Selection Method Based on Fast Condensed Nearest Neighbor Rule

Authors : Lu Si, Jie Yu, Shasha Li, Jun Ma, Lei Luo, Qingbo Wu, Yongqi Ma, Zhengji Liu

Abstract : Instance selection (IS) technique is used to reduce the data size to improve the performance of data mining methods. Recently, to process very large data set, several proposed methods divide the training set into some disjoint subsets and apply IS algorithms independently to each subset. In this paper, we analyze the limitation of these methods and give our viewpoint about how to divide and conquer in IS procedure. Then, based on fast condensed nearest neighbor (FCNN) rule, we propose a large data sets instance selection method with MapReduce framework. Besides ensuring the prediction accuracy and reduction rate, it has two desirable properties: First, it reduces the work load in the aggregation node; Second and most important, it produces the same result with the sequential version, which other parallel methods cannot achieve. We evaluate the performance of FCNN-MR on one small data set and two large data sets. The experimental results show that it is effective and practical.

Keywords : instance selection, data reduction, MapReduce, kNN
Conference Title : ICCCNC 2017 : 19th International Conference on Computer, Computing, Networking and Communications
Conference Location : Singapore, SG
Conference Dates : July 04-05, 2017