Abstract

Random Tree is a popular data classification classifier for machine learning. Feature reduction is one of the important research issues in big data. Most existing feature reduction algorithms are now faced with two challenging problems. On one hand, they have infrequently taken granular computing into thinking. On the other hand, they still cannot deal with massive data. Massive data processing is a difficult problem in the age of big data. Traditional feature reduction algorithms are generally time-consuming when facing big data. For speedily processing, we introduce a scalable fast approximate attribute reduction algorithm with Map Reduce. We divide the original data into many tiny chunks, and use reduction algorithm for each chunk. The reduction algorithm is based on correlation feature selection and generates decision rules by using Random Tree Classifier. Finally, feature reduction algorithm is proposed in data and task parallel using Hadoop Map Reduce framework with WEKA environment. Experimental results demonstrate that the proposed classifier can scale well and efficiently process big data.
2. Jiawei Han, Yanheng Liu, Xin Sun A Scalable Random Forest Algorithm Based on Map Reduce, IEEE 2013.

Index Terms

Computer Science                      Algorithms

Keywords

Hadoop, Map Reduce, Random Tree, Big Data, Correlation.