Abstract

Analysis of huge datasets has been a major concern in almost all areas of technology in the past decade and the role of data mining has become so crucial as a result of this crisis. As the data sizes in these datasets increase, from gigabytes to terabytes or even larger the complexity in collecting and warehousing these massive dataset as such in a single site is practically
impossible as it may not have enough main memory to hold all the data. Therefore they are accumulated usually in geographically distributed sites. The challenge in distributed data mining is how to learn as much knowledge from distributed databases as we do from the centralized database without costing too much communication bandwidth. A solution to distributed data mining is that the massive dataset can be collected and warehoused in a single site if its dimensionality is reduced. The dimension reduction algorithms are generally classified into feature selection, feature extraction and random projection. In this paper we propose a dimension reduction algorithm, which is different from all of these methods, to encode the transactions which reduce the size of transaction that in turn reduces the communication cost. Experimental results on a datasets demonstrate the performance of our proposed algorithm.

Reference

- Philip Machanick, A distributed systems approach to secure Internet mail, Security, Volume 24, Issue 6, September 2005, Pages 492-499
- Talia, D.“Grid-Based Distributed Data Mining Systems, Algorithms and Services, 9th International Workshop on High Performance and Distributed Mining, Bethesda April 22, 2006

Index Terms

Computer Science  
Data Mining

Key words

Centralized Database  
Data Mining

Distributed Data Mining

Dataset Dimension reduction