Abstract

In data mining, classification of data has always been an area of interest and this is especially true after the rapid increase in availability of data being collected. Cost-sensitive classification is a subset of the broader classification problem where the focus is on solving the class imbalance problem. This paper addresses the class imbalance problem using Cost-sensitive Distributed Boosting (CsDb). CsDb is a meta-classifier designed to solve the class imbalance problem for big data, is based on the concept of MapReduce. The focus of this work is to solve the class imbalance problem for the size of data which is beyond the capacity of standalone commodity hardware to handle. CsDb solves the classification problems by learning models in a distributed environment. Empirical evaluation of CsDb carried over datasets from different application domains shows average reduction of misclassification cost and number of high cost errors by 21.06% and 30.15% respectively with respect to its predecessors of type error based classifier. It preserves the cost-sensitivity of cost based predecessor. While it preserves the accuracy and F1-score, the model building time is reduced by 90.14% as compared to a non-distributed cost-sensitive classifier.
References

21. Yahoo! Webscope dataset ydata-frontpage-todaymoduleclicks\textsuperscript{v1 0} [http://labs.yahoo.com/Academic Relations].

Index Terms

Computer Science

Distributed Systems

Keywords

Class imbalance problem, distributed boosting, distributed classification