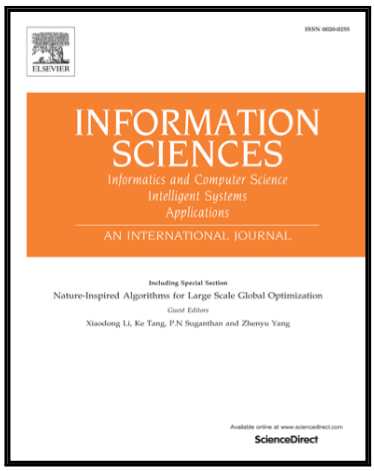

Large-Scale Distributed Sparse Class-Imbalance Learning

Chandresh Kumar Maurya, Durga Toshniwal

PII: S0020-0255(18)30359-1
DOI: 10.1016/j.ins.2018.05.004
Reference: INS 13632

To appear in: *Information Sciences*

Received date: 28 September 2017
Revised date: 1 May 2018
Accepted date: 2 May 2018

Please cite this article as: Chandresh Kumar Maurya, Durga Toshniwal, Large-Scale Distributed Sparse Class-Imbalance Learning, *Information Sciences* (2018), doi: 10.1016/j.ins.2018.05.004

Available online at www.sciencedirect.com

ScienceDirect

Information Sciences 00 (2018) 1–11

Procedia Computer Science

Large-Scale Distributed Sparse Class-Imbalance Learning

Chandresh Kumar Maurya, Durga Toshniwal[a]

[a]*Department of Computer Science & Engineering, Indian Institute of Technology, Roorkee-247667, Haridwar, U.K., India*

Abstract

Class-imbalance learning is a classic problem in data mining and machine learning community. In class-imbalance learning, the idea is to learn the model so that it performs equally well on all the classes. Most of the work in literature so far have tackled this problem either in a centralized way or the work is limited to a particular domain such as intrusion detection. In the present paper, we propose to solve the class-imbalance learning problem on *large-scale sparse* data in a *distributed setting*. More specifically, we partition the data across examples and distribute each chunk of the data to different processing nodes. Each node runs a local copy of FISTA-like algorithm which is a distributed implementation of the prox-linear algorithm for cost-sensitive learning. We show the efficacy of the proposed approach on benchmark and real-world data sets and compare the performance with the state-of-the-art techniques in the literature.

Keywords: Class imbalance learning; Distributed learning; Anomaly detection

1. Introduction

Class-imbalance learning problem is a classic problem and has been studied in the data-mining and machine learning community [38, 27, 2, 15, 8, 22]. It aims to classify each example correctly despite the imbalance proportion of each class in the data set. In the case of class-imbalance in binary classification, the goal is to correctly classify both the classes even if the proportion of classes is severely imbalanced. For example, In ad-click prediction problem, the number of ads that get clicks is much less than the number of un-clicked ads . Similarly, in the spam-email classification problem, spam-email constitutes a tiny amount of total emails. Due to severe class imbalance, traditional classification techniques fail to capture the imbalance where the focus is on correctly classifying the minority class. We emphasize here that the class-imbalance problem is similar to, on a high level, anomaly detection, outlier detection, novelty detection problems in the literature (please see the wiki page for anomaly detection), and hence, techniques devised for one problem can be applied to the other.

Current literatures handle the class-imbalance problem in two different ways: the first is data driven and the second is algorithm driven. Data-driven approaches include sampling based approaches while algorithm based approaches include ensemble-based approaches, one-class classification based approaches, and cost-sensitive based approaches. Sampling-based approaches [3, 18] try to oversample/undersample data such that different classes appear in equal proportion in the training data. The major drawback of sampling-based approaches is that they may lose important information or overfit the training data in the case of undersampling and oversampling respectively. Secondly, they are not scalable to high dimensions [15]. Ensemble-based and one-class classification based approaches suffer from high training time [13]. The reason is that ensemble-based approaches require training of a n number of models and the training time becomes a bottleneck when heterogeneous models are being trained on. The one-class classification based approaches such as one-class SVM [24], SVDD [33] work on kernel-matrices that are kept in memory and working on huge kernel matrices is a bottleneck. Recently, cost-sensitive learning has been proved to be effective in dealing with the class-imbalance in the large-scale setting [34, 37, 6, 17, 26]. In cost-sensitive learning, the idea is to give a high penalty for misclassification of the minority class so as to counterbalance the classification of each class [6]. Recently, it is proved that the cost-sensitive learning based approaches are more scalable than sampling-based approaches [40] on big data. Therefore, we choose cost-sensitive learning based approaches to handling the class-imbalance in the current work.

Most of the work in the literature mainly focus on solving the class-imbalance problem in a centralized way under the assumption that the data is available in a central repository. However, due to massive data growth in recent years, security, cost, and distributed work environment, collecting data at a central location is in-feasible. Therefore, there is a need to devise an alternative mechanism that can process data in a *distributed* way. It will not only help in decreasing the cost but also prevent the