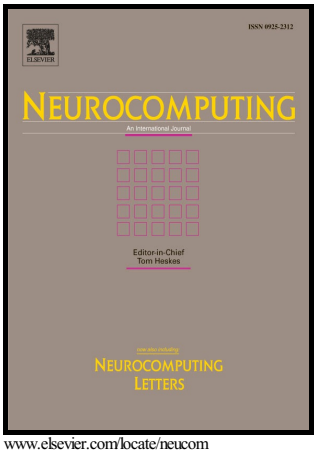

www.elsevier.com/locate/neucom

Online sparse class imbalance learning on big data

Chandresh Kumar Maurya, Durga Toshniwal, Gopalan Vijendran Venkoparao

PII: S0925-2312(16)30796-2
DOI: http://dx.doi.org/10.1016/j.neucom.2016.07.040
Reference: NEUCOM17397

To appear in: *Neurocomputing*

Received date: 17 November 2015
Revised date: 15 May 2016
Accepted date: 25 July 2016

Cite this article as: Chandresh Kumar Maurya, Durga Toshniwal and Gopalan Vijendran Venkoparao, Online sparse class imbalance learning on big data, *Neurocomputing,* http://dx.doi.org/10.1016/j.neucom.2016.07.040

Available online at www.sciencedirect.com

ScienceDirect

Neurocomputing 00 (2016) 1–12

Procedia Computer Science

Online Sparse Class Imbalance Learning on Big Data

Chandresh Kumar Maurya, Durga Toshniwal[a], Gopalan Vijendran Venkoparao[b]

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

Abstract

Class imbalance learning is the study of problems in which some classes appear more frequently than the others. Most existing works that study this problem assume data set to be dense and do not exploit the rich structure of the data. One such structure is the sparsity. In the present work, we focus on solving the class imbalance problem under the sparsity assumption. More specifically, a well-known *Gmean* metric for class imbalance learning problem in binary classification setting has been maximized, which results in a non-convex loss function. Convex relaxation techniques are used to convert the non-convex problem to the convex problem. The problem formulation in the present work uses L_1 regularized proximal learning framework and is solved via accelerated-stochastic-proximal gradient descent algorithm. Our aim in the paper is to show: (i) The application of proximal algorithms to solve real world problems (class imbalance) (ii) How it scales to Big data and (iii) How it outperforms some recently proposed algorithms in terms of *Gmean, F-measure* and *Mistake rate* on several benchmark data sets.

Keywords: Class imbalance learning; Online learning; Proximal algorithm; Big data

1. Introduction

Class imbalance learning problem is the study of classification problems in which some classes appear more frequently than others. For example, in malicious URL detection task, the number of malicious URLs detected is far less as compared to benign URLs. Similarly, in intrusion detection task, the number of suspicious users is far less as compared to normal users. A data set is said to be class-imbalanced if it contains samples from different classes in varying proportions. There have been many studies on class imbalance problem (see [1] and references therein). However, most of these works tackle the said problem in offline settings [2, 3, 4, 5]. In recent years, there have been massive amount of data being generated everywhere ranging from power plants to video surveillance systems. The data generated from such systems is high dimensional, streaming, sparse and often class imbalanced. There is a need to process streaming data in an online fashion in order to make a real-time decision.

In an effort to address the problem of class imbalance learning, researchers have used different techniques: (i) Under/Over sampling [6] (ii) Cost-sensitive learning [7] (iii) Kernel-based Methods [3]. In under/over sampling, either the majority class is under-sampled or the minority class is over-sampled in binary classification setting. The idea is to balance the distribution of two classes by sampling representative proportions of the class examples. In cost-sensitive learning, different classes are assigned different mis-classification costs because in the real world a false negative may prove costlier than a false positive. For example, denying a loan to a valuable customer is more harmful than to a fraudulent customer, from the 1business point of view. In some empirical studies and application domain, it has been shown that cost-sensitive learning produces better results than sampling-based methods [8]. Kernel-based methods use some kernel function within the kernel-based classifiers, e.g., SVM. The advantage of using kernel function is that it can find non-linear decision boundary that encompasses most of the examples from majority class, leaving aside the minority examples.

However, there are serious issues in using the aforementioned techniques. Firstly, sampling-based techniques are neither scalable to the number of samples nor to the data dimensionality in the case of Big data. Secondly, existing works on sampling such as [1, 9] do not exploit the rich structure present in the data such as sparsity. Kernel-based methods suffer from scalability and long training time. For example, if data dimensionality is of the order of millions, kernel-based methods will require storing gram matrix (also known as kernel matrix) of size million $\times$ million which is prohibitive for machines with low memory. Cost-sensitive learning has recently gained popularity in addressing the class imbalance problem [10, 11] because of: (i) learning cost to be assigned to different classes in a data dependent way (ii) scalability (iii) ability to exploit spar-