Class Imbalanced Data: Open Issues and Future Research Directions

G. Rekha ; Amit Kumar Tyagi ; N Sreenath ; Shashvi Mishra All Authors

Abstract

Document Sections

I. Introduction

II. Literature Survey

III. Motivation

IV. Issues Related to Class Imbalance Problem

V. Challenges in Solving Class Imbalance Problem

Show Full Outline

Authors

References

Keywords

Metrics

More Like This

Abstract: Since last two decades, imbalanced data is becoming a hot topic to do research or to determine meaningful results. One of the problems of machine learning and data mining... **View more**

Metadata

Abstract:

Since last two decades, imbalanced data is becoming a hot topic to do research or to determine meaningful results. One of the problems of machine learning and data mining areas is class imbalance. Data sets with imbalances have hindered the efficiency of algorithms for data mining and machine learning (in terms of overall accuracy, decision making). In the big data era, the expansion of data mining and machine learning has raised new challenges with the nature of data. In class imbalanced data, majority class lead to problem, i.e., having an imbalance between minority and majority class samples created several problems for researchers. In result, researchers are unable to learn much from systems or they are unable to find or determine prediction or take decision for respective applications like fraud detection, rare diseases identification/ prediction, approval of credit card, software defect prediction, etc. A survey for class imbalance problem is proposed in this paper with discussing several applications (where this problem getting attention). For solving this famous problem or balance this imbalanced data, three methods like Data-level, algorithm-level and hybrid methods are being considered/ used. Among these methods, a hybrid method is receiving much popularity. This paper also discusses several open issues and challenges (which are required to be developed in near future for efficient/ imbalanced learning). Also, in last several (essential) future research directions have been also discussed in this work, which makes this work as important one for research community.

Published in: 2021 International Conference on Computer Communication and Informatics (ICCCI)

Date of Conference: 27-29 Jan. 2021 **INSPEC Accession Number:** 20632183

Date Added to IEEE *Xplore*: 21 April 2021 **DOI:** 10.1109/ICCCI50826.2021.9402272

ISBN Information: **Publisher:** IEEE

Print on Demand(PoD) ISSN: 2329-7190 **Conference Location:** Coimbatore, India

Contents

I. Introduction

In classification, when the number of samples in at least one class is much greater than the samples in other classes, the question of class imbalance arises. The class with a greater number of samples will be represented as majority class and the remaining classes are called as minority class. In computer networks, the patterns related to unusual behaviour in anomaly detection of network access occur very rarely (minority class) and become the focus of interest [1]. For a class with a greater number of samples, conventional machine learning algorithms can achieve good prediction accuracy, but the performance will be degraded for class with small number of samples [2]. However, the minority class is sometimes more interesting. For example, in a medical diagnosis of a rare disease where such a rare medical disorder among ordinary populations is desperately required to be detected. Any diagnosis errors will cause patients to be at risk and have more complications. The doctors could not afford an incorrect diagnosis as this could seriously affect the well-being of the patients and even alter the course of therapies and medicines available. Basically, a class imbalance problem can be considered into two types, i.e., Binary classification problems and Multi-class imbalanced problem. It is possible to consider binary classification problems as the most evolved branch of learning from imbalanced data. It comes from different real-life applications, such as computer protection (valid operation vs. unauthorised or deceptive one), or computer vision (target object vs. background), medicine (sick vs. healthy). On the other hand, the imbalanced multi-class classification is not a well-developed binary class / counterpart (majority or not well balanced, minority or well-balanced) are no longer existed (and clearly explained).

Authors

References

Keywords

Metrics