# Chapter 14
# An Experimental Analysis to Learn Data Imbalance in Scholarly Data:
## A Case Study on ResearchGate

**Mitali Desai**
https://orcid.org/0000-0002-3264-6143
*Sardar Vallabhbhai National Institute of Technology, Surat, India*

**Rupa G. Mehta**
*Sardar Vallabhbhai National Institute of Technology, Surat, India*

**Dipti P. Rana**
https://orcid.org/0000-0002-5058-1355
*Sardar Vallabhbhai National Institute of Technology, Surat, India*

## ABSTRACT

*Data imbalance is a key challenge in the majority of real-world classification problems. It refers to the disparity of data instances corresponding to either of the class labels. Data imbalance is studied in detail with respect to many data domains such as transaction data, medical data, e-commerce data, meteorological data, social media data, and web data. But the scholarly data domain is yet to be analyzed pertaining to data imbalance. In this chapter, the scholarly data domain is explored with a focus to study various forms of data imbalance. A well-known and popular scholarly platform, ResearchGate (RG), is targeted to extract real scholarly data. An extensive experimental analysis is performed on the extracted data in order to identify the existence of both data-level and network-level imbalance. The outcome contributes to the learning of various types of data imbalance that exist in scholarly data. Resolving the existing data imbalance will substantially help in achieving efficient and accurate outcomes in many real-world scholarly literature applications.*

DOI: 10.4018/978-1-7998-7371-6.ch014

## INTRODUCTION

Data imbalance is referred as an unequal distribution of data instances among classes (Kaur et al. 2019). A few data instances fall into the minority class whereas majority class is occupied by a major portion of data instances. Such substantial difference in data distribution leads to a performance bias in the model. It is a crucial issue in many data-intensive applications such as anomaly detection, fraud discovery, spam recognition, natural disaster prediction, image recognition, disease identification and claim prediction (Somasundaram & Reddy, 2016). An intensive work has been carried out to study data imbalance in mentioned data domains.

The proposed solutions to resolve data imbalance majorly fall int two categories: data centric approaches and algorithm centric approaches. Various data centric approaches (Rout et al., 2018; Kotsiantis et al., 2006; Mahmood, 2015) include under sampling, over sampling, cluster-based over sampling, Synthetic Minority Over-sampling Technique (SMOTE). Algorithm centric approaches incorporate bagging based methods, boosting based methods, penalized algorithms and tree-based techniques to address the imbalance issue.

In recent times, applications such as community detection, influence identification, influence ranking, expert finding, recommendation systems and discovering topical authority in the domain of scholarly literature analytic are seeking researchers' interest. The open accessibility, availability and digitization provide researchers a common medium in terms of scholarly platforms to connect, communicate and cooperate. Due to an extensive utilization of various scholarly platforms, there exist massive scholarly data. Such scholarly data that is present across wide range of digitized scholarly platforms provides potential base in scholarly literature analytic.

Numerous studies are present in the realm of the mentioned applications, although identifying the existence and the forms of data imbalance in scholarly data is yet to be focused. Data imbalance leads to many adverse effects that decrease the efficiency of scholarly outcomes. On the other hand, identifying and then resolving various data imbalance will help achieving outperforming results in scholarly literature applications.

In this chapter, the real scholarly data extracted from very popular scholarly platform ResearchGate (RG) is thoroughly analyzed to inspect the data imbalance problem. In data extraction, $N$ number of RG users are targeted and their profile demographics are rendered. Based upon selected demographics, various experimentation scenarios are built to deeply analyze the data imbalance. Further, the RG graph is constructed using the followers/followings relations present among targeted RG users. The RG graph is visualized in order to identify network-level imbalance.

The results disclose that due inherent complex characteristics of scholarly data, various forms of imbalance exist at both data and network level. Identifying such forms of imbalance opens new paradigms of understanding data imbalance in order to develop novel approaches and mechanisms to achieve accurate outcomes for real-world problems in scholarly literature domain.

The remainder of this chapter is systematized as follows: In section Related Research, the recent research that is carried out in different application domains to identify and resolve data imbalance is briefly discussed. The experimental setup and results are described in section Experimentation and Result Analysis. In section Conclusion, this work is concluded.