Chapter 1
Review of Imbalanced Data Classification and Approaches Relating to Real-Time Applications

Anjali S. More
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

In today's era, multifarious data mining applications deal with leading challenges of handling imbalanced data classification and its impact on performance metrics. There is the presence of skewed data distribution in an ample range of existent time applications which engrossed the attention of researchers. Fraud detection in finance, disease diagnosis in medical applications, oil spill detection, pilfering in electricity, anomaly detection and intrusion detection in security, and other real-time applications constitute uneven data distribution. Data imbalance affects classification performance metrics and upturns the error rate. These leading challenges prompted researchers to investigate imbalanced data applications and related machine learning approaches. The intent of this research work is to review a wide variety of imbalanced data applications of skewed data distribution as binary class data unevenness and multiclass data disproportion, the problem encounters, the variety of approaches to resolve the data imbalance, and possible open research areas.

DOI: 10.4018/978-1-7998-7371-6.ch001

Copyright © 2021, IGI Global. Copying or distributing in print or electronic forms without written permission of IGI Global is prohibited.

INTRODUCTION

Data cataloging into specific classes is one of the foremost techniques in the domain of machine learning and mining with the heuristics of balanced dataset i.e. the data is equally distributed among the classes. This heuristic is not true in the existent world applications and the majority of the related applications are having imbalanced dataset where data is skewed towards one class or more than one classes. The imbalanced nature of data is having their own importance, one cannot neglect them. Thus, many researchers are motivated to deal with imbalanced classification for real-life applications. There is an incessant growth of instances of data availability in many application eras such as finance, health care, computer network system, security, internet of things, etc. where it is very much essential to advance the primary perceptive of knowledge discovery and data analysis to take the critical decision.

Nowadays, though there is existence of data discovery techniques, imbalanced data applications relating to real-life scenarios have shown the great attraction to the researchers to focus on imbalanced applications and review the problems occurred due to data unevenness. The individuals working in industry as well as academia gets attracted towards diverted data applications as review in the survey section by Alberto Fernández et al. (2009).

Several realistic application areas deal with the handling of uneven data representation, the minority instance class gets ignored due to the majority instance class. Unequal data distribution leans performance metrics towards the majority class. The review study in this research focuses on the most important application categories of imbalanced data distribution as binary class imbalance and multiclass data imbalance. To deal with the promising issues arising from class imbalance this study presents a review of imbalanced data applications, imbalanced data categories, problems encountered due to this characteristic, and the methodologies to deal with distorted data relating to real-life applications.

BACKGROUND

Classification is the most popular technique to correctly classify an instance with unknown class. Many real-world data sets show evidence of unequal class distributions in which maximum data samples are belonging to one of the larger class and far fewer data instances are falling into minority class. In case of medical diagnosis example, which consist of the cases that relates to diagnosis for a rare disease. For the referred example, only 2% of the patients are positive diagnosis and 98% diagnosis as negative. Dealing with such imbalanced datasets and related classification generates the need of machine learning algorithms. In current time the data diverted applications are relating to binary as well as multiclass data imbalance. In both category of imbalanced class either of one class having maximum instance and which diverts the performance towards majority class, i.e. performance is leaned towards majority class. The traditional classifiers reveal accurate forecast for the majority instance class and diversify the performance in case of minority data sample class. The cost of misclassification an imbalanced class can be harmful for the real world application like disease diagnostics. Thus, in today's era, Stefan Lessmann (2014) and Rebeen A. H., Masashi K. & Jens L (2020) show imbalanced data applications have received considerable attention from the research community to further boost their performance by numerous machine learning algorithms. Lars W. Jochumsen et al. (2016), Nahit Emanet et al. (2014), explained in the study that there are diverse approaches to tackle the trouble of extremely imbalanced data applications. In particular, the study deals with the description of preprocessing, cost-sensitive learning, Support Vector

20 more pages are available in the full version of this document, which may be purchased using the "Add to Cart" button on the publisher's webpage:
www.igi-global.com/chapter/review-of-imbalanced-data-classification-and-approaches-relating-to-real-time-applications/280908

Related Content

Heart Disease Prediction Using Decision Tree and Random Forest Classification Techniques
Nitika Kapoorand Parminder Singh (2021). *Applications of Big Data in Large- and Small-Scale Systems (pp. 234-259).*
www.irma-international.org/chapter/heart-disease-prediction-using-decision-tree-and-random-forest-classification-techniques/273931

Correlation-Based Ranking for Large-Scale Video Concept Retrieval
Lin Linand Mei-Ling Shyu (2010). *International Journal of Multimedia Data Engineering and Management (pp. 60-74).*
www.irma-international.org/article/correlation-based-ranking-large-scale/49150

Disease Prediction System Using Image Processing and Machine Learning in COVID-19
Sonal Raju Shilimkar, Varsha Pimpraleand Chhaya R. Gosavi (2022). *Designing User Interfaces With a Data Science Approach (pp. 134-155).*
www.irma-international.org/chapter/disease-prediction-system-using-image-processing-and-machine-learning-in-covid-19/299751

Advanced Machine Learning Algorithms for Personalized Diabetic Foot Ulcer Treatment
Madan Mohan Tito Ayyalasomayajula (2025). *AI and the Revival of Big Data (pp. 347-372).*
www.irma-international.org/chapter/advanced-machine-learning-algorithms-for-personalized-diabetic-foot-ulcer-treatment/369507

Requirements to a Search Engine for Semantic Multimedia Content
Lydia Weiland, Felix Hanserand Ansgar Scherp (2014). *International Journal of Multimedia Data Engineering and Management (pp. 53-65).*
www.irma-international.org/article/requirements-to-a-search-engine-for-semantic-multimedia-content/120126