Artificial Intelligence Review (2024) 57:137
https://doi.org/10.1007/s10462-024-10759-6

A survey on imbalanced learning: latest research, applications and future directions

Wuxing Chen[1,2] · Kaixiang Yang[3] · Zhiwen Yu[3] · Yifan Shi[4] · C. L. Philip Chen[3]

Accepted: 7 April 2024

Abstract
Imbalanced learning constitutes one of the most formidable challenges within data mining and machine learning. Despite continuous research advancement over the past decades, learning from data with an imbalanced class distribution remains a compelling research area. Imbalanced class distributions commonly constrain the practical utility of machine learning and even deep learning models in tangible applications. Numerous recent studies have made substantial progress in the field of imbalanced learning, deepening our understanding of its nature while concurrently unearthing new challenges. Given the field's rapid evolution, this paper aims to encapsulate the recent breakthroughs in imbalanced learning by providing an in-depth review of extant strategies to confront this issue. Unlike most surveys that primarily address classification tasks in machine learning, we also delve into techniques addressing regression tasks and facets of deep long-tail learning. Furthermore, we explore real-world applications of imbalanced learning, devising a broad spectrum of research applications from management science to engineering, and lastly, discuss newly-emerging issues and challenges necessitating further exploration in the realm of imbalanced learning.

✉ Kaixiang Yang
yangkx@scut.edu.cn

Wuxing Chen
ftchenwuxing@mail.scut.edu.cn

Zhiwen Yu
zhwyu@scut.edu.cn

Yifan Shi
shiyifan@hqu.edu.cn

C. L. Philip Chen
philipchen@scut.edu.cn

[1] School of Future Technology, South China University of Technology, Panyu district, Guangzhou 511442, Guangdong, China

[2] Peng Cheng Laboratory, Nanshan district, Shenzhen 518066, Guangdong, China

[3] School of Computer Science and Engineering, South China University of Technology, Panyu district, Guangzhou 510006, Guangdong, China

[4] College of Engineering, Huaqiao University, Fengze District, Quanzhou 362021, Fujian, China

Published online: 09 May 2024

Keywords Imbalanced learning · Ensemble learning · Multiclass imbalanced learning · Machine learning · Imbalance regression · Long-tailed learning

1 Introduction

In the field of machine learning, it is commonly assumed that the number of samples in each class under study is roughly equal. However, in real-life scenarios, due to practical applications in enterprises or industries, the generated data often exhibits imbalanced distribution. In the case of fault detection, for example, the major class has a large number of samples, while the other class (the class with faults) has only a small number of samples (Ren et al. 2023). As depicted in Fig. 1, this imbalance presents a challenge as traditional learning algorithms tend to favor the more prevalent classes, potentially overlooking the less frequent ones. Nevertheless, from a data mining perspective, minority classes often carry valuable knowledge, making them crucial. Consequently, the objective of imbalanced learning is to develop intelligent systems capable of effectively addressing this bias, thereby enabling learning algorithms to handle imbalanced data more effectively.

Over the past two decades, imbalanced learning has garnered extensive research and discussion. Numerous methods have been proposed to tackle imbalanced data, encompassing data pre-processing, modification of existing classifiers, and algorithmic parameter tuning. The issue of imbalanced data classification is prevalent in real-world applications, such as fault detection (Fan et al. 2021; Kuang et al. 2021; Ren et al. 2023), fraud detection, medical diagnosis (Hung et al. 2022; Fotouhi et al. 2019; Behrad and Abadeh 2022), and other fields (Yang et al. 2022; Haixiang et al. 2017; Ding et al. 2021). In these application scenarios, datasets typically exhibit a significant class imbalance, where a few classes contain only a limited number of samples, while the majority class is more abundant. This imbalance leads to varying performance of learning algorithms, known as performance bias, on the majority and minority classes. To effectively address this challenge, imbalanced learning has been

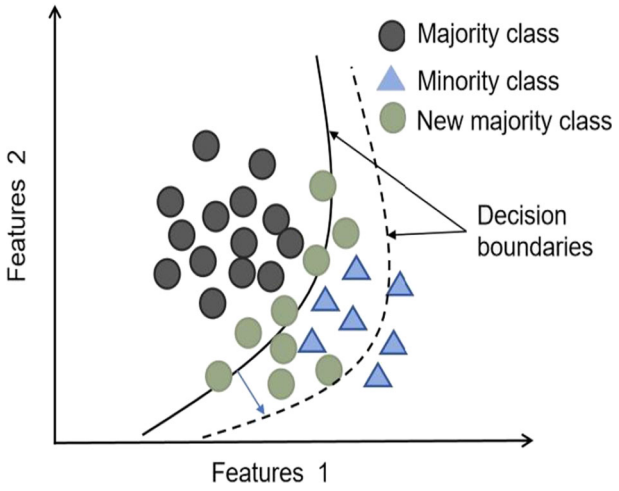

Fig. 1 A binary imbalanced dataset. The decision boundaries learned on such samples tend to be increasingly biased towards the majority class, leading to the neglect of minority class data samples. The dashed line is the decision boundary after the addition of the majority class. Note that the new samples are added arbitrarily, just to show how the added samples affect the decision boundaries

extensively examined by both academia and industry, resulting in the proposal of various methods and techniques (Chen et al. 2022a, b; Sun et al. 2022; Zhang et al. 2019).

The processing of imbalanced data has gained significant importance due to the escalating volume of data originating from intricate, large-scale, and networked systems in domains such as security (Yang et al. 2022), finance (Wu and Meng 2016) and the Internet (Di Mauro et al. 2021). Despite some notable achievements thus far, there remains a lack of systematic research that reviews and discusses recent progress, emerging challenges, and future directions in imbalanced learning. To bridge this gap, our objective is to present a comprehensive survey of recent research on imbalanced learning. Such an endeavor is crucial for sustaining focused and in-depth exploration of imbalanced learning, facilitating the discovery of more effective solutions, and advancing the field of machine learning and data mining.

In this paper, we categorize the existing innovative solutions for imbalanced data classification algorithms into five types, as illustrated in Fig. 2. These types encompass general methods, ensemble learning methods, imbalanced regression and clustering, long-tail learning, and imbalanced data streams. Within these categories, we further delve into more detailed methods, including data-level approaches, algorithm-level techniques, hybrid methods, general ensemble frameworks, boosting, bagging, cost-sensitive ensembles, imbalanced regression, imbalanced clustering, online or ensemble learning, concept drift, incremental learning, class rebalancing, information enhancement, and model improvement. Employing this classification scheme, we conduct a comprehensive review of existing imbalanced learning approaches and outline recent practical application directions.

Table 1 summarizes the differences between recent reviews on imbalanced learning and the present investigation. Distinguishing this survey from previous works, we not only summarize recent solutions to the category imbalance problem in traditional machine learning but also address the long-tailed distribution issue in deep learning, which has gained prominence. Additionally, we extensively elaborate on the imbalance problem within the unsupervised or semi-supervised domain. Moreover, building upon current research, we not only encapsulate emerging solutions in imbalanced learning but also outline new challenges and future research

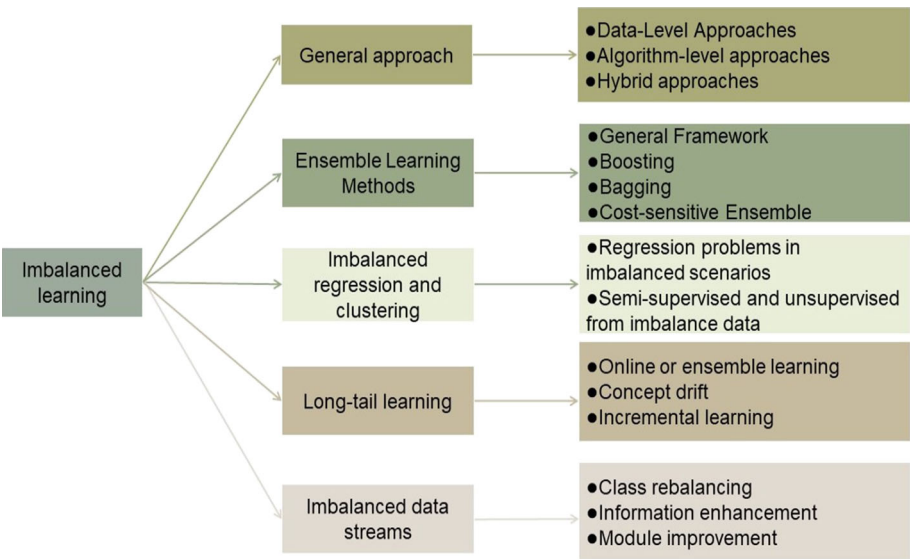

Fig. 2 Classification of existing methods

Table 1 Overview of different investigations on imbalanced learning

Title	Years	Categories	Task Description
Knowledge discovery from imbalanced and noisy data (Van Hulse and Khoshgoftaar 2009)	2009	Binary classification	Review sampling algorithms and solutions from noisy imbalanced data
Learning from imbalanced data (He and Garcia 2009)	2009	Binary classification, regression	Review sampling methods, cost-sensitive learning and methods based on kernel learning and active learning
A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches (Galar et al. 2012)	2012	Binary classification	Review sampling methods, ensemble learning algorithms (bagging, boosting, etc.), and hybrid algorithms
Class Imbalance Problem in Data Mining: Review (Longadge and Dongre 2013)	2013	Binary classification	Review data-level methods, algorithm-level methods, and feature selection class methods
A survey of multiple classifier systems as hybrid systems (Wózniak et al. 2014)	2014	Binary classification, application research	Review ensemble learning algorithms, hybrid algorithms, and intelligent hybrid systems for multiple classifier systems
Learning from imbalanced data: open challenges and future directions (Krawczyk 2016)	2016	Binary classification, multi-classification, application research	Review of sampling method, cost-sensitive learning method, ensemble learning, application fields and future research directions

Table 1 continued

Title	Years	Categories	Task Description
A broad review on class imbalanced learning techniques (Rezvani and Wang 2023)	2016	Binary classification, regression, application research	Definition of classification problem, nature of problem, general learning algorithm, application research field, Imbalanced algorithms based on support vector machines
Learning from class-imbalanced data: Review of methods and applications (Haixiang et al. 2017)	2017	Binary classification, regression, multi-classification, application research	Review sampling methods, cost-sensitive learning methods, ensemble learning, and application areas
A Systematic Review on Imbalanced Data Challenges in Machine Learning: Applications and Solutions (Kaur et al. 2019)	2019	Binary classification, classifier introduction, application research	Review applied research, data preprocessing methods, cost-sensitive learning methods, hybrid methods, and classifiers
Deep Long-Tailed Learning: A Survey (Zhang et al. 2023)	2023	Deep long-tail learning, computer vision, multi-class classification	Review of deep imbalanced classification methods for long-tailed distributions, class rebalancing, information augmentation, and model boosting
OURS	/	Binary classification, multi-classification, **regression**, application exploration, **clustering**, data stream, **deep long-tail learning**	Reviews the data-level, algorithm-level, hybrid methods, ensemble learning, **deep long-tail learning**, **regression**, **clustering and data streaming algorithms**

directions with practical potential. The organisational framework of this paper is shown in Fig. 3.

We summarize the main contributions of the survey in this paper as follows:

- **This paper provides a unified and comprehensive review of imbalanced learning and deep imbalanced learning.** This paper presents the inaugural comprehensive review and summary of the existing research outcomes in imbalanced learning and deep imbalanced learning. It systematically consolidates a wide range of methods and techniques, thereby facilitating researchers in developing a comprehensive understanding of this field.
- **A comprehensive survey of long tail learning and imbalanced machine learning applications.** This paper presents a comprehensive survey and summary of the applications of long-tail learning and imbalanced machine learning over the past few years, encompassing diverse fields and real-world application scenarios. Through this study, scholars may get a more profound comprehension of the use of imbalanced learning in several fields, so functioning as a beneficial resource for further inquiries.
- **Six new research challenges and directions were identified.** Drawing upon the existing research findings, this paper identifies and proposes six novel research challenges and directions within the realm of imbalanced learning. These challenges and directions hold significant potential and research importance, contributing to the advancement and progression of the imbalanced learning field.

The remainder of this paper is organized as follows: Section 2 outlines the current research methodology employed in this study and presents preliminary statistics on imbalanced learn-

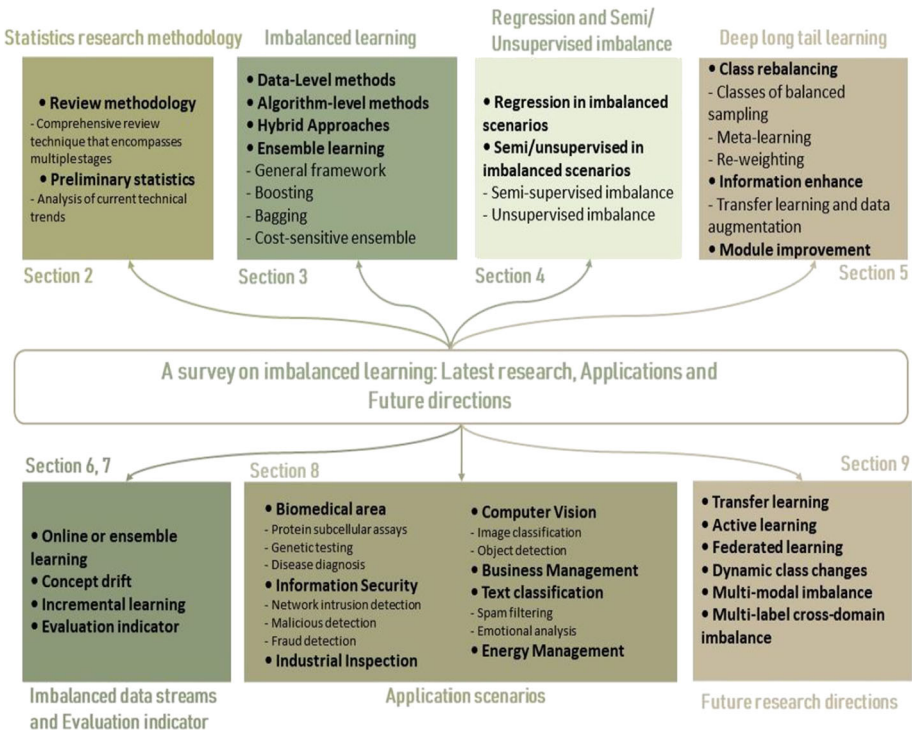

Fig. 3 The organizational framework of this paper

ing. Section 3 introduces a comprehensive approach for addressing imbalanced equilibrium learning. Section 4 delves into the methodologies associated with imbalanced regression and clustering. Section 5 explores methods specifically pertaining to long tail learning. Section 6 discusses relevant research on imbalanced data streams. In Section 8, we categorize existing research on imbalanced learning applications into seven fields and provide an overview of the respective research contents within each field. Section 9 compiles and summarizes six future research directions and challenges in imbalanced learning based on the survey and current research trends. Finally, we conclude this paper with a concise summary.

2 Statistics research methodology

2.1 Review methodology

The statistical classification research method proposed in this paper draws upon the works (Kaur et al. 2019; Haixiang et al. 2017). This work uses a multi-stage thorough review strategy to capture a wide range of methods and real-world application domains related to imbalanced learning.

In the initial stage, we focus on imbalanced, sampling, and skewed data to obtain preliminary results. Subsequently, in the second stage, we conduct a thorough search of existing research literature. For the third stage, we employ a triple keyword search approach. Initially, we use data mining, machine learning, and classification keywords to assess the research status of machine learning technologies. Next, we incorporate long-tail distribution and neural network as keywords to examine the research status of deep long-tail learning. Finally, we employ keywords such as ”detection,” ”abnormal,” or ”medical” to identify practical applications in the literature.

To ensure comprehensive coverage, we conducted searches across seven databases encompassing various domains within the natural and social sciences. These databases include IEEE Xplore, ScienceDirect, Springer, ACM Digital Library, Wiley Interscience, HPCsage, and Taylor & Francis Online.

2.2 Preliminary statistics

Figure 4 illustrates the search framework employed in this study. Initially, we applied the search terms ”imbalance” or ”unbalance” to filter the initial research works based on these keywords. Subsequently, we employed a trinomial tree search strategy to filter abstracts, conclusions, and full papers, leading to the identification of papers falling into three distinct directions. To ensure comprehensive research coverage, we incorporated the keywords ”machine learning,” ”long-tail learning”, and ”applications” during the trinomial tree screening stage. Following this, we conducted manual screening to eliminate duplicates and obtain the final results, resulting in a total of 2287 papers. This systematic search process ensures an extensive exploration of the field of imbalanced learning, providing a robust support and foundation for this paper.

Figure 5 demonstrates the publication trend of papers in the field of imbalanced learning from January 2013 to July 2023. The number of published papers remained relatively stable until 2017, with a slight decline observed during the 2015-2016 period. However, there has been a remarkable surge in the number of published papers from 2018 onwards. This trend indicates the enduring significance of imbalanced learning as a research topic, with ongoing growth in research hotspots. Furthermore, we compiled a list of the top 20 journals or conferences contributing to paper publications. As depicted in Fig. 6, prominent publications

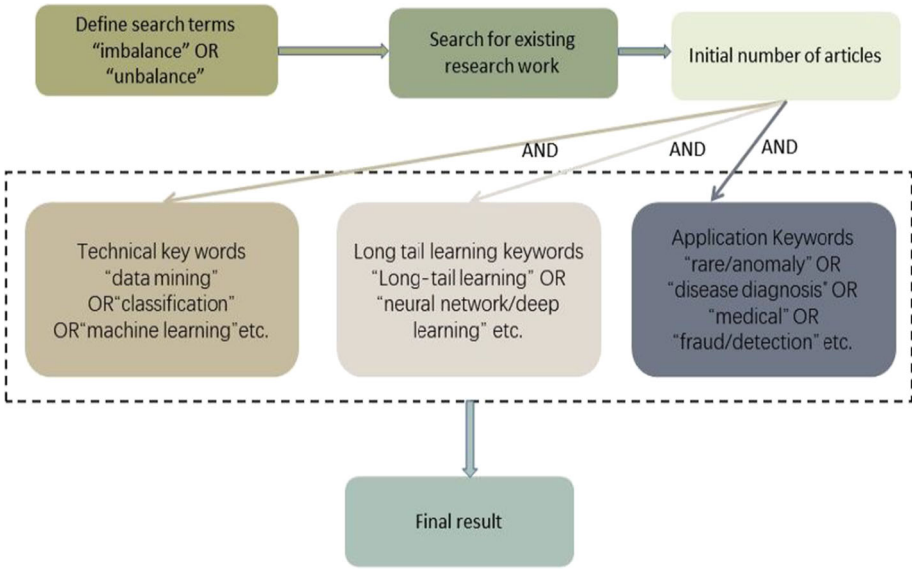

Fig. 4 Search framework for obtaining papers

are found in fields such as computer science, neural computing, management science, and energy applications.

3 Imbalanced data classification approaches

In this section, we first introduce data-level approaches in imbalanced data classification, then we give various solutions at the algorithmic level as well as hybrid approaches, and finally we introduce more robust ensemble learning methods.

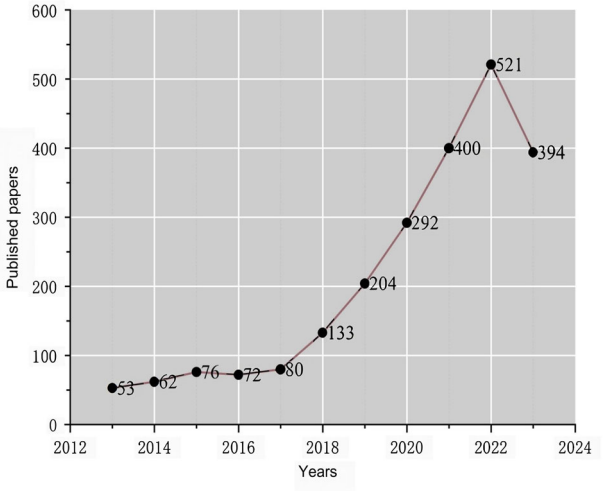

Fig. 5 Trends in imbalanced Learning Dissertation Publications

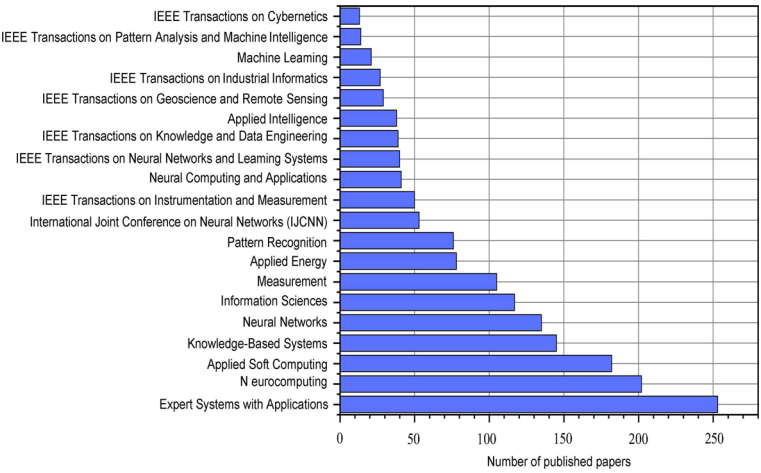

Fig. 6 Top 20 journals or conferences that publish papers on imbalanced learning

3.1 Data-Level approaches

Data-level approaches to the problem of class imbalance involve techniques and methods that directly modify the training data. These approaches target the challenge of class imbalance by rebalancing the distribution of classes within the training dataset. This rebalancing can be achieved through undersampling, which involves reducing instances from the majority class, or oversampling, which involves increasing instances from the minority class. The objective is to create a more balanced dataset that can effectively train a binary classification model.

The most widely used oversampling methods are random replication of a few samples and synthetic minority over-sampling technique (SMOTE) (Chawla et al. 2002). However, SMOTE is prone to generate duplicate samples, ignore sample distributions, introduce false samples, are not suitable for high-dimensional data, and are sensitive to noise. Therefore, there are many variants of methods that improve on SMOTE. Borderline-SMOTE (Han et al. 2005) identifies and synthesizes samples near the decision boundary, ADASYN (He et al. 2008) adapts the generation of synthetic samples based on minority class density, and Safe-Level-SMOTE (Bunkhumpornpat et al. 2009) incorporates a safe-level algorithm to balance class distribution while reducing misclassification risks. Other methods do not rely solely on synthetic sampling techniques. They explore different strategies for the selection of preprocessing stages or categorization algorithms to suit the characteristics and needs of imbalanced datasets. Stefanowski and Wilk (2008) introduced a novel method for selective pre-processing of imbalanced data that combines local over-sampling of the minority class with filtering challenging cases from the majority classes. To enhance the classification performance of SVM on unbalanced datasets, Akbani et al. (2004) offered several methodologies and procedures, such as altering class weights, utilizing alternative kernel functions, adjusting decision thresholds.

It is worth noting that oversampling methods sometimes produce redundant noise problems, so some oversampling sampling methods combine noise processing in the sampling process to obtain a more excellent balanced dataset. Stefanowski et al. introduced SMOTE-IPF, an extension of SMOTE that incorporates an iterative ensemble-based noise filter called IPF, effectively addressing the challenges posed by noisy and borderline examples in imbalanced datasets (Sáez et al. 2015). In order to handle class imbalance in a noise-free way,

Douzas et al. Douzas et al. (2018) introduced a straightforward and efficient oversampling methodology called k-means SMOTE, which combines k-means clustering and the synthetic minority oversampling technique (SMOTE). Sun et al. (2022) introduced a disjuncts-robust oversampling (DROS) method that effectively addresses the challenge of negative oversampling results in the presence of small disjuncts, by utilizing light-cone structures to generate new synthetic samples in minority class areas, outperforming existing oversampling methods in handling class imbalance. At present, there are some new innovative data sampling methods that can combine the regional distribution of data and random walk based on data stochastic mapping to synthesize new samples, and even map data to higher-dimensional space to generate artificial data points (Sun et al. 2020; Zhang and Li 2014; Douzas and Bacao 2017; Han et al. 2023).

Undersampling is the practice of decreasing the number of examples in the majority class in order to obtain a more balanced class distribution. Undersampling seeks to match the number of instances in the minority class with the number of instances in the majority class by randomly or purposefully deleting samples from the majority class. The most widely used is undersampling using KNN technology (Wilson 1972; Mani and Zhang 2003). Wilson (1972) studied the asymptotic characteristics of the nearest neighbor rule using edited data, and discusses the influence of edited data methods on classification accuracy. Mani and Zhang (2003) proposed a method for processing imbalanced dataset distribution based on k nearest neighbor algorithm by taking information extraction as a case study. Michael and colleagues examine cases that are often misclassified and look at elements that contribute to their difficulty in order to better comprehend machine learning data and develop training procedures and algorithms (Smith et al. 2014).

In recent years, some researchers have been investigating variants of sampling methods that take into account the density information, spatial information, and intrinsic characteristics between classes of the data and combine them with reinforcement learning, deep learning, and clustering algorithms for data enhancement. For example, Kang et al. introduced WU-SVM, an improved algorithm that utilizes a weighted undersampling scheme based on space geometry distance to assign different weights to majority samples in subregions, allowing for better retention of the original data distribution information in each learning iteration (Kang et al. 2017). Yan et al. proposed Spatial Distribution-based UnderSampling (SDUS), an undersampling method for imbalanced learning that maintains the underlying distribution characteristics by utilizing sphere neighborhoods and employing sample selection strategies from different perspectives (Yan et al. 2023). Fan et al. proposed a sample selection strategy based on deep reinforcement learning algorithm, aiming to optimize the sample distribution, by constructing a reward function, using deep reinforcement learning algorithm to train the model to select more representative samples, thereby improving the diagnostic accuracy (Fan et al. 2021). Kang et al. (2016) introduced a new under-sampling scheme that incorporates a noise filter before resampling, resulting in improved performance of four popular under-sampling methods for imbalanced classification. Other researchers have incorporated clustering algorithms into undersampling. In order to successfully handle class imbalance issues, Tsai et al. (Lin et al. 2017) presented a unique undersampling strategy called cluster-based instance selection (CBIS), which combines clustering analysis with instance selection. Han et al. (2023) proposed a combination of global and local oversampling methods to distinguish between minority and majority classes by comparing them to the discretisation values at each class level. Instances are synthesised based on the degree of discrete magnitude.

The oversampling method offers several advantages in addressing imbalanced data, including increased representation of the minority class, retention of original information, improved classifier performance, and reduced bias towards the majority class. By artificially increasing

the number of instances in the minority class, the dataset becomes more balanced for training, allowing classifiers to better learn the characteristics and patterns of the minority class. However, oversampling can also have drawbacks, such as the potential for overfitting, increased computational complexity, the introduction of noise into the dataset, and limited information gain. These disadvantages can impact the classifier's generalization ability and performance on unseen data. When using the oversampling approach, it is necessary to take the specific oversampling methodology and the features of the dataset into account. A comparison of algorithmic studies of the latest relevant data levels in recent years is shown in Table 2.

Sampling methods have been of interest to researchers mainly because they are described according to the following guidelines: (1) they work independently of the learning algorithm, (2) they lead to a balanced redistribution of data, and (3) they can be easily combined with any learning mechanism. Of course different sampling methods can be categorised based on the complexity of the corresponding sampling method. Under-sampling typically has faster run times and higher robustness and is less prone to overfitting than oversampling.

3.2 Algorithm-level approaches

Algorithm-level approaches concentrate on adapting or developing machine learning algorithms to effectively handle imbalanced datasets. These techniques prioritize enhancing the algorithms' capability to accurately classify instances from the minority class. By adjusting or designing algorithms to be more responsive to these underrepresented groups, these techniques contribute to improving the overall performance, fairness, and generalizability of machine learning models when confronted with imbalanced data. Among the most prevalent algorithm-level strategies is cost-sensitive learning, wherein the classification performance is enhanced by modifying the algorithm's objective function. This alteration ensures that the model receives greater emphasis on learning from underrepresented classes (Castro and Braga 2013; Yang et al. 2021). Table 3 shows the advantages and disadvantages of the main algorithm-level approaches in recent years.

3.2.1 Cost-sensitive learning

Cost-sensitive learning takes into account the different costs associated with different types of misclassification and aims to optimise the model for scenarios where the consequences of errors are uneven. Lately, researchers have incorporated other techniques based on cost-sensitive learning to enable models with higher classification accuracy and better generalization. For example, Zhang and Hu (2014) proposed cost-free learning (CFL) as a method for achieving optimal classification outcomes without depending on cost information, even when class imbalance exists. CFL approach maximizes normalized mutual information between targets and decision outputs, enabling binary or multi-class classifications with or without abstaining. To enhance the classification performance of support vector machines (SVMs) on imbalanced data sets, Cao et al. (2020). developed the unique approach ATEC. By assessing its performance in terms of classification accuracy and changing it in the proper direction, ATEC effectively optimizes the error cost for between-class samples. Addressing both concept drift and class imbalance problems in streaming data, Lu et al. (2019) proposed an adaptive chunk-based dynamic weighted majority technique. In addressing the imbalanced class problem, Fan et al. (2017) introduced Entropy-based Fuzzy Support Vector Machine (EFSVM) to addresses the imbalanced class problem by assigning fuzzy memberships to samples based on their class certainty, resulting in improved classification

Table 2 Comparison of typical state-of-the-art methods at the data level

Algorithm	Advantages	Disadvantages	Years
Disjuncts-robust oversampling (Sun et al. 2022)	Tackling the challenge of small disjuncts in oversampling	Inability to solve multi-class imbalance problems and handle large data sets	2022
Spatial distribution under-sampling (Yan et al. 2023)	Selection of a representative majority sample	Inability to adequately fit the entire distribution of the original unclassified data	2023
Sampling combining clustering and instance selection (Tsai et al. 2019)	Improving the performance of ensemble classifiers	Data-dependent and high computational complexity	2019
Global local-based oversampling (Han et al. 2023)	Intra-class variation of data and selective generation strategies are considered	Data dependency, model complexity	2023
Sampling strategies with reinforcement learning (Fan et al. 2021)	Autonomous sample selection to improve classification performance	Complex algorithms and reliance on data feedback	2021

Table 3 Comparison of typical state-of-the-art methods at the algorithm level

Algorithm	Advantages	Disadvantages	Years
Error cost auto-tuning SVM (Cao et al. 2020)	Improved F1-score and AUC in a minority class	Inability to solve multi-class imbalance problems and handle large data sets	2021
Adaptive Weighted ensemble BLS (Yang et al. 2021)	Simple, efficient and low-complexity	Failure to address multi-class imbalance problem	2021
Adaptive chunk-based dynamic weighted majority (Lu et al. 2019)	Addressing dynamic imbalance problem	Complex models with long run times	2020
Kernel-based class-specific BLS (Chen et al. 2022a)	Ability to handle imbalanced multi-class problems with noise	Not suitable for large-scale datasets	2022
Class-specific weighted RVFLN (Sahani and Dash 2019)	Low computational complexity and high noise immunity	Vulnerable to data distribution	2019

performance compared to other algorithms. Datta and Das (2019, 2015) proposed another idea where a multi-objective optimisation framework (Datta et al. 2017) was used to train SVM to efficiently find the best trade-off between objectives in class imbalanced data sets. Cao et al. (2021) proposed an adaptive error cost adjustment method for class imbalance learning of support vector machines (SVMs) called ATEC, which has a significant advantage over the traditional grid search strategy in terms of training time by efficiently and automatically adjusting the error cost between samples. This novel domain can effectively solve imbalance problems without expensive parameter tuning.

3.2.2 Weighted shallow neural networks

Shallow neural networks combined with imbalanced learning algorithms have made extensive developments in solving imbalanced data classification problems. Shallow neural networks have fast training capability. Due to the characteristics of the shallow layer structure, the training time of neural networks is relatively short. This is very beneficial for applications that quickly process imbalanced datasets and get instant results. For example, In order to accurately and efficiently identify and categorize power quality disturbances, Sahani and Dash (2019) proposed an FPGA-based online power quality disturbances monitoring system that makes use of the reduced-sample Hilbert-Huang Transform (HHT) and class-specific weighted Random Vector Functional Link Network (RVFLN). Choudhary and Shukla (2021) proposed an integrated ELM method for decomposing complex imbalance problems into simpler subproblems for the class bias problem in classification tasks. The technique successfully addresses elements like class overlap and the number of probability distributions present in the unbalanced classification issue by utilizing cost-sensitive classifiers and cluster assessment. Chen et al. proposed a novel approach called double-kernelized weighted broad learning system (DKWBLS) that addresses the challenges of class imbalance and parameter tuning in broad learning systems (BLS). By utilizing a double-kernel mapping strategy, DKWBLS generates robust features without the need for adjusting the number of nodes. Additionally, DKWBLS explicitly considers imbalance problems and achieves improved decision boundaries (Chen et al. 2022b). Chen et al. proposed a new double kernel-based class-specific generalized learning system (DKCSBLS) to solve the multiclass imbalanced learning problem. DKCSBLS solves the imbalance multiclassification problem based on class distribution adaptively combined with class-specific penalty coefficients and uses a dual kernel mapping mechanism to extract more robust features (Chen et al. 2022a). Yang et al. addresses the imbalance problem in the broad learning system (BLS) by proposing a weighted BLS and an adaptive weighted BLS (AWBLS) that consider the prior distribution of the data. Additionally, an incremental weighted ensemble broad learning system (IWEB) is proposed to enhance the stability and robustness of AWBLS (Yang et al. 2021).

Algorithmic-level approaches optimize the loss function associated with the dataset, concentrating on a limited number of classes to enhance model performance. In contrast to resampling-based methods, these techniques are more computationally efficient and better suited for large data streams. Consequently, considering their ability to enhance AUC and G-mean in imbalanced scenarios along with runtime considerations, algorithmic-level methods may be preferred. Furthermore, they offer flexibility in selecting distinct activation functions and optimization algorithms tailored to specific problem requirements.

These approaches aim to address imbalance classification problems by adapting existing algorithms or creating new ones specifically designed for imbalanced datasets, emphasizing a smaller number of classes. They are easy to implement as they do not necessitate dataset

or feature space modifications and are applicable across a wide range of classification algorithms, potentially enhancing classifier performance on unbalanced datasets. However, they may not fully capture the complexity of imbalanced datasets and can be sensitive to hyperparameter selection. Additionally, they do not openly tackle data scarcity among minority groups, and their efficiency may vary based on unique dataset features.

3.3 Hybrid approaches

To address imbalanced learning, hybrid methods incorporate techniques from both the data-level and the algorithm-level. In an effort to improve performance, these techniques try to combine the advantages of both strategies. Table 4 shows the advantages and disadvantages of the main hybrid methods in recent years.

A common hybrid approach is to combine sampling methods with ensemble learning (Sağlam and Cengiz 2022; Abedin et al. 2022; Razavi-Far et al. 2019; Sun et al. 2018; Liang et al. 2022). Another is the combination of sampling methods using multiple optimisation algorithms or adaptive domain methods. For example, Sağlam and Cengiz (2022) proposed a method to solve the category imbalance problem in classification called SMOTEWB (SMOTE with boosting), which combines a new noise detection method with SMOYE. By combining noise detection and augmentation in an ensemble algorithm, SMOTEWB overcomes the challenges associated with random oversampling (ROS) and SMOTE by adjusting the number of neighbours per observation in the SMOTE algorithm. Abedin et al. (2022) presented WSMOTE-ensemble, a unique ensemble technique for small company credit risk assessment that addresses the issue of severely uneven default and nondefault classes. To build robust and varied synthetic examples, the suggested technique combines WSMOTE and Bagging with sampling composite mixes, thereby eliminating class-skewed limitations. Razavi-Far et al. (2019) presented unique class-imbalanced learning algorithms that combine oversampling methods with bagging and boosting ensembles, with a particular emphasis on two oversampling strategies based on single and multiple imputation methods. To rebalance the datasets for training ensemble algorithms, the suggested strategies attempt to develop synthetic minority class samples with missing values estimate that are similar to the actual minority class samples. For the evaluation of unbalanced business credit, Sun et al. (2018) developed the DTE-SBD decision tree ensemble model, which combines the SMOTE with the Bagging ensemble learning algorithm with differential sample rates.

Another is the combination of sampling methods using multiple optimisation algorithms or adaptive domain methods. Chen et al. (Pan et al. 2020) proposed Gaussian oversampling and Adaptive SMOTE methods. Adaptive SMOTE takes the Inner and Danger data from the minority class and combines them to form a new minority class that improves the distributional features of the original data. Gaussian oversampling is a technique that combines dimensionality reduction with a Gaussian distribution to make the distribution's tails narrower. Dixit and Mani (2023) proposed a novel oversampling filter-based method called SMOTE-TLNN-DEPSO, a hybrid variant of the method that combines SMOTE for synthetic sample generation and the Differential Evolution-based Particle Swarm Optimisation (DEPSO) algorithm for iterative attribute optimisation. Yang et al. (2019) proposed a hybrid optimum ensemble classifier framework that combines density-based under-sampling and cost-effective techniques to overcome the drawbacks of traditional imbalanced learning algorithms.

In order to overcome the cost problems associated with parameter optimisation of traditional hybrid methods. datta et al. (Mullick et al. 2018, 2019) combined neural networks

Table 4 Comparison of typical state-of-the-art methods at the hybrid level

Algorithm	Advantages	Disadvantages	Years
Boosting-SMOTE (Sağlam and Cengiz 2022)	Ability to detect noise in imbalanced data	Poor adaptability and inability to solve multi-classification problems	2022
Filtering-based oversampling methods (Dixit and Mani 2023)	Effective resolution of noise and boundary samples	Consumes more runtime	2023
Sampling strategies for single and multiple interpolation (Razavi-Far et al. 2019)	The diversity of base classifiers is considered	Failure to address multiclass imbalance problem	2019
Entropy-based hybrid sampling (Li et al. 2019)	Consider differences in information content between classes	Complex models with long run times	2019
Adaptive SMOTE gaussian sampling (Pan et al. 2020)	More effective in categorising extremely imbalanced data	More complex models and longer run times	2020

and heuristic algorithms with sampling methods or traditional classifiers. For example, the performance of the k-nearest neighbour classifier on imbalanced data sets was improved by adjusting the k value using neural networks or heuristic learning algorithms. Another is a three-way adversarial strategy that combines convex generators, multi-class classifier networks and discriminators in order to perform oversampling in deep learning systems.

Hybrid methods can integrate various imbalance learning techniques, such as combining clustering algorithms with sampling methods and coupling them with cost-sensitive algorithms to enhance model performance. However, this integration can lead to increased model complexity and necessitate more parameter tuning. Therefore, when selecting hybrid methods to address imbalance algorithms, it is essential to consider their runtime and model complexity.

3.4 Ensemble learning methods

Ensemble learning is a methodology employed for the classification of imbalanced data, which involves amalgamating multiple classifiers or models to enhance the performance of classification tasks on datasets characterized by imbalanced class distributions. Its primary objective is to tackle the challenges presented by imbalanced class distributions by capitalizing on the strengths of different classifiers, thereby enhancing the predictive accuracy for the minority classes (Yang et al. 2021). Ensemble methods for imbalanced data classification typically encompass the creation of diverse subsets from the imbalanced dataset through resampling techniques. Individual classifiers are then trained on these subsets, and their predictions are combined using voting or weighted averaging schemes. Ensemble learning efficiently reduces the bias towards the majority class and enhances overall classification performance on imbalanced data sets by integrating different models.

Over the past few years, many scholars have summarised the progress, challenges and potential solutions of ensemble learning on the problem of unbalanced data classification. At the same time, mainstream methods are categorised and discussed, challenges are clarified, and research directions are proposed. Finally, many surveys have also explored the possibility of combining ensemble learning with other machine learning techniques (Dong et al. 2020; Yang et al. 2023; Galar et al. 2012). For example, Galar et al. (2012) explored the challenges posed by imbalanced datasets in classifier learning. It reviews integrated learning methods for the problem of classifying unbalanced data, proposes classification criteria, and makes empirical comparisons. It is demonstrated that random undersampling combined with either bagging or boosting ensembles works well, emphasizing the superiority of ensemble-based algorithms over stand-alone preprocessing methods. Table 5 summarizes the advantages and disadvantages of some main ensemble learning algorithms over the last three years.

3.4.1 General framework

General ensemble is a generalisability framework for addressing imbalanced learning that is robust to multiple datasets. The goal behind generic integration is to use the diversity and complementary qualities of various models to improve the integration's overall performance and resilience. This approach aims to overcome the limitations of a single model by aggregating their predictions through various techniques such as voting, averaging or weighted combinations. Liu et al. (2020) proposed to generate strong integrations for unbalanced classification through self-paced coordination of under-sampled data hardness to cope with the challenges posed by imbalanced and low-quality datasets. This computationally efficient method takes

Table 5 Comparison of typical state-of-the-art methods with ensemble learning

Algorithm	Advantages	Disadvantages	Years
Self-paced Ensemble (Liu et al. 2020)	Ability to cope with noise and large data sets	Ensemble learning frameworks are complex	2022
Boosting and bagging ensemble with imputation (Razavi-Far et al. 2021)	Increased sampling diversity for greater robustness	High computational complexity	2021
ECUBoost (Wang et al. 2020)	Combining entropy and confidence to avoid information loss	Long calculation time	2020
Double evolution bagging framework (Guo et al. 2022)	More focus on classifier performance and ensemble combinations	Inability to guarantee diversity of base classifiers	2022
Subspace sampling ensemble (Xu et al. 2021)	Capable of handling high-dimensional imbalances	The sampling method is relatively simple	2021

into account class disproportionality, noise and class overlap to achieve robust performance even in highly skewed distributions and overlapping classes, while being applicable to a variety of existing learning methods. Liu et al. (2020) present a novel integrated imbalanced learning framework called MESA, which adaptively resamples the training set to create multiple classifiers and form a cascaded integrated model. MESA learns the sampling strategy directly from the data, going beyond heuristic assumptions to optimize the final metric. Liu et al. (2009) proposed two algorithms, EasyEnsemble and BalanceCascade, to address the undersampling in the class imbalance problem. easyEnsemble combines the output of multiple learners trained on a majority class subset, while BalanceCascade sequentially removes the majority class examples that are correctly classified in each step.

The general framework of ensemble learning offers several advantages in addressing the classification of imbalanced data. Firstly, it provides a flexible and adaptable approach that can incorporate various ensemble methods, such as bagging, boosting, or cost-sensitive techniques, depending on the specific characteristics of the imbalanced dataset. This versatility allows for customization and optimization based on the specific problem at hand. Additionally, ensemble learning can effectively leverage the diversity of multiple classifiers to improve the overall classification performance and handle the challenges posed by imbalanced data, such as class overlap or rare class identification. It can also provide robustness against noise or outliers in the data.

However, there are also noteworthy potential disadvantages to consider when employing ensemble learning approaches. Firstly, such approaches may necessitate additional computational resources and increased training time in comparison to single classifier methods. This is due to the requirement of training and combining multiple models within the ensemble. Secondly, the performance of ensemble learning is highly dependent on several factors, including the selection of appropriate base classifiers, ensuring their diversity, and employing an effective ensemble combination strategy. These factors demand careful consideration and tuning to achieve optimal results. Furthermore, the adaptability of ensemble learning to specific domains is limited. While generic framework models may perform well in certain datasets or domains, their efficacy may be compromised in others. As a result, precise tweaking and optimization of the ensemble models may be necessary to achieve high performance in a variety of settings.

3.4.2 Boosting

Boosting ensemble learning is a machine learning technique that combines multiple weak classifier iteratively to create a strong ensemble model. The most widely used of these is the Adaboost (Freund and Schapire 1997), smoteboost (Chawla et al. 2003) and RUSBoost (Seiffert et al. 2009). Every weak classifier is trained on a subset of the data, prioritizing the samples that were misclassified by the preceding models. This iterative process aims to improve the overall performance of the ensemble by focusing on challenging instances during training. The final prediction is obtained by aggregating the predictions of all weak learners, typically using weighted voting or averaging. Boosting effectively improves the overall performance and generalization of the ensemble by emphasizing the challenging instances and reducing the bias in the learning process.

In the past several years, researchers have devoted much thought and attention to Boosting techniques for solving the class imbalance problem, aiming to improve the classification accuracy of minority class and to improve the learning performance in the case of class imbalance. For example, Kim et al. (2015) proposed GMBoost, which deals with data imbalances based on geometric means. GMBoost can comprehensively consider the learning of majority

and minority classes by using the geometric mean of the two classes in error rate and accuracy calculations. Roozbeh et al. et al. (Razavi-Far et al. 2021) proposed a new class imbalanced learning technique that combines oversampling methods with bagging and boosting ensembles. The article proposes two strategies based on single and multiple imputation to create synthetic minority class samples by estimating missing values in the original minority class to improve classification performance on imbalanced data. Wang et al. (2020) proposed the ECUBoost framework, which combines augmented integration-based with novel entropy and confidence-based undersampling methods to maintain the validity and structural distribution of the majority of samples during undersampling to address the imbalance problem. In recent years Datta et al. (2020) proposed a new Boosted method that achieves the trade-off between majority and minority classes without expensive search costs. The method treats the weight assignment of component classifiers as a game of tug-of-war between classes in the edge space and avoids costly cost-set adjustments by implementing efficient class compromises in a two-stage linear programming framework.

In summary, Boosting has the following advantages in solving the classification of imbalanced data. (1) It can effectively deal with class imbalance by focusing on a small number of classes and assigning higher weights to misclassified instances, thereby improving overall classification performance. (2) Boosting can combine multiple weak classifiers to create strong integration, resulting in better generalisation and robustness. They can adaptively adjust the weights of classifiers during boosting iterations, emphasising difficult instances and reducing bias towards majority classes.

However, it is important to acknowledge the limitations and potential drawbacks of the Boosting algorithm. One notable concern is its relatively higher computational cost, particularly when handling large-scale datasets. The iterative nature of the enhancement process necessitates multiple iterations and the training of weak classifiers, which can significantly increase training time and resource requirements. Furthermore, augmentation algorithms are sensitive to the presence of noisy or mislabeled data, which can have a detrimental impact on their performance. It is crucial to address these limitations and take appropriate measures to mitigate their effects when applying the Boosting in practical settings. Furthermore, Boosting ensemble strategy is a serial iterative method that requires constant updating of the model and therefore a long training time. Therefore runtime needs to be considered when using Boosting with ensemble learning.

3.4.3 Bagging

Bagging Breiman (1996) is an ensemble learning technique that involves creating multiple subsets of the original dataset by random sampling and replacement. Each subset is used to train a separate base learner, such as a decision tree, using the same learning algorithm. The predictions from all the base learners are then combined by majority voting (for classification) or averaging (for regression) to make the final prediction. By averaging multiple independently trained models, Bagging helps to reduce variation in predictions, thereby improving generalisation and robustness (Wang and Yao 2009).

Researchers have recently concentrated on the application of bagging ensemble learning techniques to address issues with class imbalance. For example, Bader-El-Den et al. (2018) proposed a novel ensemble-based approach, called biased random forest, to address the class imbalance problem in machine learning. The technique concentrates on boosting the number of classifiers representing the minority class in the ensemble rather than oversampling the minority class in the dataset. By identifying critical areas using the nearest neighbor algorithm and generating additional random trees. Błaszczyński and Stefanowski (2015) investigated

extensions of bagging ensembles for imbalanced data, comparing under-sampling and over-sampling approaches, and proposes Neighbourhood Balanced Bagging as a new method that considers the local characteristics of the minority class distribution. Guo et al. (2022) proposed a dual evolutionary Bagging framework that combines resampling techniques and integration learning to solve the class imbalance problem. The framework aims to find the most compact and accurate integration structure by integrating different base classifiers. After selecting the best base classifiers and using an internal integration model to enhance diversity, the multimodal genetic algorithm finds the optimal combination based on mean G-means

In summary, bagging offers several advantages in addressing the classification of unbalanced data. Firstly, it can obtain a more balanced dataset and improve classification performance by random sampling. In addition, bagging reduces the variance of the model by creating multiple subsets of the original data and training multiple base classifiers independently, which helps to reduce overfitting and improve generalization. Finally, bagging is a simple and straightforward implementation that can be combined with a variety of basic classifiers. Bagging requires neither adjusting the weight update formula nor changing the amount of computation in the algorithm and is able to achieve good generalization with a simple structure. Therefore, when using ensemble learning algorithms to deal with imbalance problems, Bagging may be the better method to choose if the low complexity and running time of the model as well as the robustness of the model are important.

Although bagging can improve overall classification performance, it may still struggle to accurately classify a small number of classes if they are severely under-represented. In addition, bagging may not be effective when dealing with datasets with overlapping classes or complex decision boundaries. It is also worth noting that the performance of bagging relies heavily on the choice of the underlying classifier and the quality of the individual models in the integration.

3.4.4 Cost-sensitive ensemble

Cost-sensitive ensemble learning takes into account the imbalance costs associated with different classes and aims to optimise the overall cost of misclassification. Such methods take into account cost factors in the decision making process. By assigning appropriate weights to adjust the decision thresholds of individual classifiers, cost-sensitive ensemble techniques aim to minimise the overall cost of misclassification and to improve the performance of specific classes that bear higher costs.

Presently, an increasing number of scholars have focused on research that cost-sensitive ensemble learning. The most widely used is Adacost (Fan et al. 1999), proposed by Fan in 1996. AdaCost is a misclassification cost-sensitive boosting method that updates the training distribution based on misclassification costs, aiming to minimize cumulative misclassification costs more effectively than AdaBoost, with empirical evaluations demonstrating significant reductions in costs without additional computational overhead. Additionally, many scholars have been delving deeper into cost-sensitive ensemble in recent years from various angles. By employing a cascade of straightforward classifiers trained with a subset of AdaBoost, Viola and Jones (2001) demonstrated a unique method for quick identification in areas with highly skewed distributions, such as face detection or database retrieval. The suggested approach significantly outperforms traditional AdaBoost in face identification tasks thanks to its high detection rates, extremely low false positive rates, and quick performance. Zhang et al. (Ng et al. 2018) introduced a new incremental ensemble learning method that addresses concept drift and class imbalances in a streaming data environment, in which the class imbalances are addressed by an imbalance inversion bagging method, which is specifically applied to predict

Australia's electricity price. Akila and Reddy (2018) developed a cost-sensitive Risk Induced Bayesian Inference Bagging model, RIBIB, for detecting credit card fraud. RIBIB used a novel bagging architecture that included a limited bag formation approach, a Risk Induced Bayesian Inference base learner, and a cost-sensitive weighted voting combiner. Zhang et al. (2022) proposed an integrated framework, BPUL-BCSLR, for data-driven mineral prospectivity mapping (MPM) that addresses the challenges of imbalanced geoscience datasets and cost-sensitive classification. The proposed approach integrates Bagging-based positive-unlabeled learning (BPUL) with Bayesian cost-sensitive logistic regression (BCSLR) and was implemented for the study of MPM in the Wulong Au district, China.

Cost-sensitive ensemble learning is valuable for imbalanced datasets. It accounts for misclassification costs, vital in real-world scenarios where rare class errors are expensive. By integrating cost factors, it improves decision-making and resource allocation. These methods balance error types, enhancing sensitivity to minority classes, thus improving overall classification and class distribution representation. While cost-sensitive ensemble learning has its advantages, there are certain challenges associated with its implementation. Estimating error costs demands prior knowledge or expert input, introducing subjectivity. Selecting an effective ensemble combination strategy considering cost factors can be complex. It is not easy to optimize cost-based objectives while maintaining variety. As a result, data distribution bias often affects cost-sensitive ensemble approaches, which might result in comparatively good performance but poor resilience.

4 Regression and Semi/unsupervised learning in imbalanced data

In this section, we first describe the solution of the regression problem in an imbalance scenario. At the same time we give the solution ideas of semi-supervised and unsupervised in imbalance problems.

4.1 Regression in imbalanced scenarios

There is a notable gap in systematically exploring the imbalanced perspective of machine learning algorithms in the context of the regression problem. The regression problem in an unbalanced scenario arises when predicting continuous values, where the target variables in the dataset exhibit an imbalanced distribution. Traditional regression tasks aim to make accurate predictions for all target values. However, in imbalanced scenarios, certain target values have significantly fewer instances, resulting in skewed datasets (Krawczyk 2016; Rezvani and Wang 2023; Yang et al. 2021). This presents a challenge for regression models as the limited availability of training data may hinder their ability to accurately predict these infrequent target values.

The majority of the research on the imbalanced regression problem has focused on developing assessment metrics (Torgo and Ribeiro 2009) that consider how important observations are as well as techniques for handling undersampling and outliers in continuous output prediction. For example, Branco et al. (2017) proposed SMOGN, a novel pre-processing technique tailored for addressing imbalanced regression tasks. SMOGN addresses the performance degradation observed in rare and relevant cases in imbalanced domains. Branco et al. (2019) proposed three new methods: adapting to random oversampling, introducing Gaussian noise, and proposing a new method called WERCS (Weighted Correlation-based Combinatorial Strategy) to address the problems posed by imbalanced distributions in regression tasks.

In order to solve the problem of class imbalance in ordinal regression, Zhu et al. (2019) developed SMOR (Synthetic Minority oversampling methodology for imbalanced Ordinal Regression). SMOR takes into account the classification order and gives low selection weights to prospective generation directions that can skew the structure of the ordinal sample.

In recent years, a number of researchers have addressed the imbalance regression problem at the algorithmic level and at the ensemble learning level. For example, Branco et al. (2018) introduced the REsampled BAGGing (REBAGG) algorithm, an ensemble method designed to address imbalanced domains in regression tasks. REBAGG incorporates data pre-processing strategies and utilizes a bagging-based approach. By employing nightly pulse oximetry to diagnose obstructive sleep apnea, Gutiérrez-Tobal et al. (2021) suggested a least-squares boosting (LSBoost) model for predicting the apnea-hypopnea index (AHI). The model achieves high diagnostic performance in both community-based non-referral and clinical referral cohorts, demonstrating its ability to generalize. Kim et al. (2019) introduced a novel method for predicting river discharge (Q) utilizing hydraulic variables collected from remotely sensed data termed Ensemble Learning Regression (ELQ). The ELQ method combines multiple functions to reduce errors and outperforms traditional single-rating curve methods. Liu et al. (2022) proposed an ensemble learning assisted method for accurately predicting fuel properties based on their molecular structures. By comparing two descriptors, COMES and CM, the optimized stacking of various base learners is achieved to efficiently screen potential high energy density fuels (HEDFs) and accurately predict their properties. Steininger et al. (2021) proposed DenseWeight, a sample weighting approach based on kernel density estimation, and DenseLoss, a cost-sensitive learning approach for neural network regression. DenseLoss adjusts the influence of each data point on the loss function according to its rarity, leading to improved model performance for rare data points. Ren et al. (2022) proposed a novel loss function specifically designed for the imbalanced regression task. They offer multiple implementations of balanced MSEs, including one that does not require prior knowledge of the training label distribution. Addressing data imbalance in real-world visual regression.

In the context of regression problems, many solutions developed for categorical imbalanced data can be extended, but currently lack adequacy in addressing the challenges specific to regression imbalanced scenarios. To enhance the robustness and predictive power of regression models in unbalanced data, the utilization of integrated learning methods holds promise. By combining the prediction results from multiple regression models, it becomes possible to mitigate errors and biases associated with infrequent groups, thereby improving the overall prediction performance. The advantage lies in effectively leveraging the collective strengths of multiple models. However, caution must be exercised in controlling the diversity of integrated models to prevent overfitting.

4.2 Semi/unsupervised learning in imbalanced scenarios

Semi/unsupervised learning in imbalanced scenarios involves training machine learning models when labeled data is scarce or imbalanced across classes. These approaches leverage both labeled and unlabeled data to improve model performance, addressing challenges posed by class imbalance. Unsupervised learning method aspect involves imbalanced clustering problems, such as the case where some clusters contain more points than others. This is because traditional clustering methods may have difficulty in accurately identifying and representing clusters of a few classes leading to problems such as poor clustering and loss of information about a few cluster classes. In semi-supervised imbalanced learning (Wei et al. 2021; Yang

and Xu 2020), the challenge is not only the lack of sufficient labelled samples, but also that the distribution of these labelled samples exhibits class imbalance.

Semi-supervised imbalanced learning (SSIL) is a key problem in dealing with imbalanced data when labelled samples are scarce. For example, Chen et al. (Wei et al. 2021) proposed Class-Rebalancing Self-Training, which iteratively retrains the SSIL model and selects minority class samples more frequently. Distribution Aligning Refinery of Pseudo-label (DARP) (Kim et al. 2020) optimises the generated pseudo-labels to fit the models that are biased towards the majority class, improving the generalisation ability of SSIL under the balancing test criterion. In addition, a scalable SSIL algorithm that introduces an auxiliary balanced classifier (ABC) (Lee et al. 2021) successfully copes with class imbalance by introducing balance in the auxiliary classifier. Other studies (Yang and Xu 2020; Lee et al. 2021) have argued for the value of unbalanced labelling. Under more unlabelled data conditions, the original labels can be used for semi-supervised learning along with additional data to reduce label bias and significantly improve the final classifier performance.

To address the imbalance problem in the unsupervised case, specialised algorithms and techniques have been developed to ensure fair validity of clustering results. These methods aim to enhance the representatives of minority clusters, take into account imbalance factors, and achieve a fairer clustering distribution. For example, Nguwi and Cho (2010) combined support vector machines and ESOM for variable selection and clustering ordered features. Zhang et al. (2019), Lu et al. (2019) considered unbalanced clusters by integrating interval-type type II fuzzy local metrics. Zhang et al. (2023) proposed k-means algorithm for adaptive clustering weights, which optimised the trade-off between each cluster weight to solve the imbalanced clustering problem. Cai et al. (2022) in order to mine fused location data, developed a unique clustering technique to deal with the problem of imbalanced datasets. The OSRCIH method proposed by Wen et al. (2021) combines autonomous learning and spectral rotation clustering to tackle the challenges of imbalanced class distribution and high dimensional. These combined considerations such as variable selection, fuzzy metrics, and local density.

Overall, the methods proposed for the problem of clustering and semi-supervised learning of imbalanced data show some promise. Schemes such as automatically determining the centre of clusters and the number of clusters are well suited to arbitrarily shaped imbalanced data sets. However, unsupervised methods may require careful tuning of parameters and evaluation using specialised metrics. Although the performance of the model on imbalanced data can be significantly improved by semi-supervised learning, the correlation between unlabelled data and raw data has a significant impact on the results of semi-supervised learning, and SSIL does not really integrate a strategy for imbalanced learning, even though there is still a lot of room for improvement.

5 Deep learning classification problems under long-tail distribution

The long-tailed class distribution refers to a specific pattern observed in datasets where the occurrence of classes follows a long-tailed or power-law distribution. This distribution exhibits a small number of classes with a high frequency of instances, known as the ”head,” while the remaining classes have significantly fewer instances, forming the ”tail.” Consequently, the tail classes are commonly referred to as the minority classes, whereas the head classes are considered the majority classes. A comprehensive review of current research on deep long-tail distributions and future developments can be found in Zhang et al.’s literature

(Zhang et al. 2023). This distribution pattern is frequently encountered in various real-world scenarios, including image detection (Zang et al. 2021), visual relation learning (Desai et al. 2021), and few-shot learning (Wang et al. 2020), where certain classes are more prevalent than others. Ghosh et al. (2022) demonstrate through many experiments that the category imbalance problem is not eliminated by deep learning, and also provide many solutions that have been offered so far to solve this problem, which are generally categorised into post-processing, pre-processing, and dedicated algorithms.

Long-tail learning and class imbalanced learning are two interrelated yet distinct research areas. Long-tail learning can be viewed as a specialized sub-task of class imbalanced learning, with the main distinction being that in long-tail learning, the samples of tail classes are typically very sparse and do not necessarily exhibit an absolute imbalance in the number of classes. In contrast, class imbalanced learning typically involves some minority class samples (Zhang et al. 2023; Li et al. 2021). Despite these differences, both research areas are dedicated to addressing the challenges posed by class imbalance and share certain ideas and approaches, such as class rebalancing, when developing advanced solutions. At the same time, Ghosh et al. (2022) explored whether the effect of class imbalance on deep learning models is related to its effect on their shallow learning counterparts, with the aim of exploring the effect of class imbalance on deep learning models and whether deep learning has fully solved the problem in machine learning.

5.1 Class rebalancing

Recent research on long-tail distributions can be divided into the following three categories: class rebalancing, information enhancement, and module improvement (Zhang et al. 2023). Class rebalancing is one of the main approaches for long-tail learning, which aims to address the negative effects of class imbalance by rebalancing the number of training samples. Recent deep long-tail research has used various classes of balanced sampling methods, rather than random resampling, for the training of small batches of deep models. However, these strategies require prior knowledge of the frequency of training samples for different categories, which may not be available in practice (Zhang et al. 2023; Kang et al. 2019).

New research work has recently proposed that rebalancing any imbalanced categorical dataset should essentially just rebalance the classifier, and should not change the distribution of picture features for feature learning with the distribution of categories (Zhou et al. 2020). Wang et al. (2019) propose a unified framework called Dynamic Course Learning (DCL), which adaptively adjusts the sampling strategy and weights in each batch. DCL combines a two-level course scheduler for data distribution and learning importance, resulting in improved generalization and discriminatory power. Zhang et al. (2021) proposed FrameStack, a frame-level sampling method that dynamically balances the class distribution during training, thereby improving video recognition performance without compromising overall accuracy.

There are also methods that incorporate meta-learning (Liu et al. 2020; Hospedales et al. 2021). Zang et al. (2021) propose a method called Feature Augmentation and Sampling Adaptive (FASA) to address the challenge of data scarcity for rare object classes in long-tail instance segmentation. FASA uses an adaptive feature augmentation and sampling strategy to augment the demand space for rare classes, using information from past iterations and adjusting the sampling process to prevent overfitting. By adaptively balancing the impact of meta-learning and task-specific learning within each task, Lee et al. (2019) introduced Bayesian Task-Adaptive Meta-Learning (Bayesian TAML), a revolutionary meta-learning

model that tackles the drawbacks of previous techniques. By learning the balancing variables, Bayesian TAML determines whether to rely on meta-knowledge or task-specific learning for obtaining solutions. Dablain et al. (2022) proposed DeepSMOTE, an oversampling algorithm for deep learning models, which addresses the challenge of unbalanced data by combining an encoder/decoder framework, SMOTE-based oversampling, and a penalty-enhanced loss function.

Another approach is the re-weighting related research method, which solves the long-tail or imbalanced distribution problem by improving the loss, which is often simple to implement and requires only a few lines of code to modify the loss to achieve a very competitive result. Cui et al. (2019) proposed a novel theoretical framework for addressing the problem of long-tailed data distribution by measuring data overlap using small neighboring regions instead of single points. In order to establish class balance in the loss function, a reweighting method is created using the effective number of samples, which is determined depending on the volume of data. Muhammad et al. (Jamal et al. 2020) proposed a meta-learning method to explicitly estimate the differences between class conditional distributions, which enhances classical class balancing learning by linking class balancing methods to domain adaptation. Cao et al. (2019) proposed label distribution-aware margin (LDAM) loss as well as delayed re-weighting training schemes that minimise margin-based generalisation bounds and allow the model to learn the initial representation before applying re-weighting, thus improving the performance of imbalanced learning.

Class rebalancing is a simple but well-performing method in long-tail learning, especially when inspired by class-sensitive learning. This makes it an attractive option for real-world applications. However, class rebalancing methods are usually performance trade-offs, and improving the performance of the tail classes may decrease the performance of the head classes. To overcome this problem, combining different approaches can be considered, but the pipeline needs to be carefully designed to avoid performance degradation (Zhang et al. 2023). This suggests that the long-tail problem may require more information-enhanced approaches to effectively deal with tail class deficiencies.

5.2 Information enhancement

To increase the performance of deep learning models in long-tailed learning scenarios, a strategy known as the Information Enhancement Method is used in relation to deep long-tail distribution. It is centered on enriching the available information during model training. Transfer learning and data augmentation are the two primary areas that this approach covers (Zhang et al. 2023).

Transfer learning aims to learn generic knowledge from the head common class and then transfer it to the tail less sample class. Recently, there has been a growing interest in applying migration learning to scenarios with deep long-tail distributions. For example, Liu et al. addressed the challenge of learning deep features from long-tailed data by proposing a method that expands the distribution of tail classes in the feature space. The method augments each instance of tail classes with disturbances, creating a ”feature cloud” that provides higher intra-class variation. With this method, deep representation learning on long-tailed data is enhanced since it reduces the distortion of the feature space brought on by the unequal distribution between head and tail classes. Xiang et al. (2020) introduced a novel framework for self-paced knowledge distillation. This approach includes two levels of adaptive learning schedules, namely self-paced expert selection and lesson example selection, aiming to effectively transfer knowledge from multiple ’experts’ to a unified student model. Wang et al.

(2020) addressed the challenge of imbalanced classification in long-tailed data by proposing a new classifier called RoutIng Diverse Experts (RIDE). To lower model variance, decrease model bias, and lower computing costs, RIDE makes use of many experts, a distribution-aware diversity loss, and a dynamic expert routing module. Further research has found that experimental results indicate that self-supervised learning plays a positive role in learning a balanced feature space for long-tailed data (Kang et al. 2020). Furthermore, research is being done to investigate methods for managing long-tailed data with noisy labels (Karthik et al. 2021).

Data augmentation in the context of deep long-tailed distributions refers to a technique that aims to improve the size and quality of datasets used for model training. It involves applying pre-defined transformations, such as rotation, scaling, cropping, or flipping, to each data point or feature in the dataset (Shorten and Khoshgoftaar 2019; Zhang et al. 2023). This category is divided into head-to-tail transfer enhancement and non-transfer enhancement. In head-to-tail transfer augmentation, the data augmentation process involves transferring augmented samples from the head classes to the tail classes. By applying pre-defined transformations to the samples from the head classes and adding them to the tail class data, the augmented tail class data is enriched, allowing for better representation and learning of the tail classes. This approach helps to mitigate the class imbalance issue and improve the generalization ability of the model on the tail classes. For example, Kim et al. (2020) proposed to enhance less frequent classes by performing sample translations from more frequent classes. They enable the network to learn more generalisable features for a small number of classes as a way to address the class imbalance in deep neural networks. Chen et al. (2022) proposed a reasoning-based implicit semantic data augmentation method to address the performance degradation of existing classification algorithms caused by long-tailed data distributions. By borrowing transformation directions from similar categories using covariance matrices and a knowledge graph, they generate diverse instances for tail categories. Zhang et al. (2022) proposed a data augmentation method based on Bidirectional Encoder Representation from Transformers (BERT) to address the long-tailed and imbalanced distribution problem in Mandarin Chinese polyphone disambiguation. They incorporate weighted sampling and filtering techniques to balance the data distribution and improve prediction accuracy. Dablain et al. (2023) proposed a three-stage CNN training framework with extended oversampling (EOS), aiming to address the generalisation gap of a few classes in imbalanced image data by exploiting an end-to-end training approach, learning data augmentation in the embedding space and fine-tuning the classifier head.

Information enhancement is compatible with other methods such as class rebalancing and modular improvement, especially in the two subtypes of information enhancement, migration learning and data augmentation, which, with careful design, can improve the performance of the tail categories without degrading the performance of the head categories. However, it is important to note that simply applying category-independent enhancement techniques may not be effective enough because they ignore the category imbalance problem, may add more samples from the head category than from the tail category, and may introduce additional noise. Thus how to better perform data augmentation for long-tail learning still requires further research.

5.3 Module improvement

In addition to class rebalancing and information enhancement methods, researchers have explored ways to improve the model in recent years. These methods can be divided into

representation learning, classifier design and decoupled training (Zhang et al. 2023). This method involves analyzing the challenges posed by the distribution and making targeted modifications to the model's architecture, loss function, training strategies, or data augmentation techniques to better handle the inherent biases and class imbalances.

Representation learning aim to learn feature representations that capture the underlying structure and discriminative information in the data, while also addressing the imbalance issue. For example, Chen et al. (2021) proposed a novel method based on the principles of causality, leveraging a meta-distributional scenario to enhance sample efficiency and model generalization. Liu et al. (2023) proposed Transfer Learning Classifier (TLC) to address the challenges of class-imbalanced data and real-time visual data in computer vision. The TLC model incorporates an active sampling module to dynamically adjust skewed distribution and a DenseNet module for efficient relearning. Kuang et al. (2021) proposed a class-imbalance adversarial transfer learning (CIATL) network to address the challenges of cross-domain fault diagnosis when dealing with class-imbalanced and machine faulty data. The CIATL network incorporates class-imbalanced learning and double-level adversarial transfer learning to learn domain-invariant and class-separate diagnostic knowledge. Recent studies also have explored contrastive learning approaches for addressing long-tailed problems. Methods such as KCL (Kang et al. 2020), PaCo (Cui et al. 2021), Hybrid (Wang et al. 2021), and DRO-LT (Samuel and Chechik 2021) have been proposed, each introducing innovative techniques such as k-positive contrastive loss, parametric learnable class centers, prototypical contrastive learning, and distribution robust optimization, respectively. These methods seek to reduce class imbalance, boost model generalization, and strengthen the learnt models' resistance to distribution change.

Traditional deep learning classification algorithms prioritize the majority class, resulting in poor minority class performance. And the loss functions of most classifiers are based on linear functions. To overcome these problems, various techniques have been developed to improve classifier design in the context of long-tailed distributions. In recent years, different classifier designs have been proposed to address the deep long-tail distribution problem. The Realistic Taxonomic Classifier (RTC) (Wu et al. 2020) uses hierarchical classification, mapping images into a class taxonomic tree structure. Samples are adaptively classified at different levels based on difficulty and confidence, prioritizing correct decisions at intermediate levels. The causal classifier applies a multi-head strategy to capture bias information and mitigate long-tailed bias accumulation (Tang et al. 2020). The GIST classifier (Liu et al. 2021) transfers the geometric structure of head classes to tail classes, improving performance on tail classes by enhancing tail-class weight centers through displacements from head classes. Zhou et al. (2022) proposed a unique debiased SGG approach dubbed DSDI to address the dual imbalance problem in scene graph generation (SGG). The strategy efficiently addresses the uneven distribution of both foreground-background occurrences and foreground relationship categories in SGG datasets by adding biased resistance loss and a causal intervention tree.

Decoupled training addresses this problem by dividing the training process into two stages: representation learning and classifier learning. This approach allows the model to learn a more discriminative representation of the data, effectively capturing the inherent characteristics of both the majority and minority classes. By decoupling the training, decoupled training methods have shown promising results in improving the classification accuracy and generalization of models in the context of deep long-tail distributions. Nam et al. (2023) demonstrated effectiveness in long-tail classification through separate decoupled learning of representation learning and classifier learning. The approach includes training the feature extractor using stochastic weight averaging (SWA) to obtain a generalised representation,

and a novel classifier retraining algorithm using stochastic representation and uncertainty estimation to construct robust decision bounds. Kang et al. (2020) used a k-positive contrastive loss to create a more balanced and discriminative feature space, which improved long-tailed learning performance. MiSLAS found that data mixup enhances representation learning but has a negative or negligible effect on classifier training, proposing a two-stage approach with data mixup for representation learning and label-aware smoothing for better classifier generalization.

Module improvement methods solve long-tail problems by changing network modules or objective functions. These techniques complement decoupling training and provide a conceptually simple approach to solving real-world long-tail application problems. However, such methods tend to have high computational complexity, no guaranteed substantial improvements, complex model design, lack of generalizability, and risk of overfitting. These methods require careful consideration and customization for specific scenarios.

6 Imbalanced learning in data streams

Imbalanced data streams refer to continuously arriving data instances in a streaming environment, characterized by a highly skewed and uneven class distribution. These data streams present unique challenges due to the amalgamation of streaming and imbalanced data characteristics, including concept drift and evolving class ratios. Effective mining of imbalanced data streams necessitates adaptable algorithms capable of swiftly adapting to changing decision boundaries, imbalance ratios, and class roles, while maintaining efficiency and cost-effectiveness. In the last few years, several reviews (Fernández et al. 2018; Aguiar et al. 2023; Alfhaid and Abdullah 2021) have provided comprehensive insights into the development of techniques for imbalanced data streams. These reviews offer an overview of data stream mining methods, discuss learning difficulties, explore data-level and algorithm-level approaches for handling skewed data streams, and address challenges such as emerging and disappearing classes, as well as the limited availability of ground truth in streaming scenarios. Aguiar et al. (2023) have presented a comprehensive experimental framework, evaluating the performance of 24 state-of-the-art algorithms on 515 imbalanced data streams. This framework covers various scenarios involving static and dynamic class imbalance, concept drift, and incorporates both real-world and semi-synthetic datasets. Recent methods for addressing imbalanced data streams can be classified into three categories: (1) online or ensemble learning approaches, (2) incremental learning approaches, and (3) concept drift handling methods.

6.1 Online or ensemble learning

Online or ensemble learning in solving imbalanced data streams refers to the use of algorithms and techniques that continuously update and adapt classifiers to handle the evolving and imbalanced nature of streaming data, either by incorporating multiple classifiers or by updating the classifier's model online with new incoming data, to improve classification accuracy and performance in imbalanced scenarios. For example, Du et al. (2021) introduced a cost-sensitive online ensemble learning algorithm that incorporates several equilibrium techniques, such as initializing the base classifier, dynamically calculating misclassification costs, sampling data stream samples, and determining base classifier weights. Furthermore, certain researchers have explored algorithmic techniques for online cost-sensitive learning,

integrating online ensemble algorithms with batch mode methods to address cost-sensitive bagging or boosting algorithms. Wang and Pineau (2016), Klikowski and Woźniak (2022), Jiang et al. (2022). Zyblewski et al. (2021) presented a novel framework that combines non-stationary data stream classification with data analysis of skewed class distributions, using stratified bagging, data preprocessing, and dynamic ensemble selection methods. In addition to exploring high-dimensional imbalanced data streams, recent research has also explored the problem of incomplete imbalanced data streams. You et al. (2023) proposed a novel algorithm called OLI2DS for learning from incomplete and imbalanced data streams, addressing the limitations of existing approaches. The method uses empirical risk minimisation to detect information features in the missing data space.

The combination of ensemble learning and active learning presents a powerful strategy for addressing the imbalanced data streams. It not only enhances the learning process by actively selecting informative samples but also utilizes the collective knowledge of multiple classifiers to improve classification performance. This integrated approach offers a valuable solution for applications where data arrives continuously and exhibits imbalanced characteristics. For example, Halder et al. (2023) introduced an autonomous active learning strategy (AACE-DI) for handling concept drifts in imbalanced data streams. The method incorporates a cluster-based ensemble classifier to select informative instances, minimizing expert involvement and costs. It prioritizes uncertain, representative, and minority class data using an automatically adjusting uncertainty strategy. Zhang et al. (2020) presented a novel method called Reinforcement Online Active Learning Ensemble for Drifting Imbalanced data stream (ROALE-DI). The approach addresses concept drift and class imbalance by integrating a stable classifier and a dynamic classifier group, prioritizing better performance on the minority class.

Combining ensemble learning methods with sampling methods has proved to be a very effective and popular solution for imbalanced data streams problems in recent years (Krawczyk et al. 2017; Aguiar and Cano 2023). The combination of ensemble learning and sampling both by updating or adding classifiers and in setting unique policies for differently skewed data can provide a unique solution to conceptual drift and class imbalance (Aguiar et al. 2023). Aguiar et al. (2023) apply the combination of ensemble learning methods at the data level to the field of imbalanced data flows for a detailed classification. One of the most popular and at the same time effective solutions in recent years is the combination of sampling methods with Bagging. For example, The robust online self-tuning ensemble introduced by Cano and Krawczyk (2022) addresses the challenges of concept drift, evolving class distributions, and non-smooth class imbalances by combining online training, concept drift detection, sliding windows for class-specific adaptation, and self-tuning bagging. Klikowski and Woźniak (2022) proposed deterministic sampling classifiers with weighted Bagging, which demonstrated excellent performance on a variety of imbalance ratios, label noise levels, and conceptual drift types through data preprocessing and weighted Bagging. On the other hand, the effectiveness of the integration approach in dealing with unbalanced data streams can be further enhanced by employing specialised combinatorial strategies or block-based adaptive learning (Aguiar et al. 2023). For example, Yan et al. (2022) proposed a dynamically weighted selection integration that dynamically adjusts the attenuation factor of the base classifier by resampling a small number of samples from previous data blocks. Feng et al. (2022) proposed an incremental learning algorithm, DME, which uses distribution matching and adaptive weighting integration to efficiently deal with concept drift in real-world streaming datasets.

Online learning approaches have lower runtime and model complexity in solving imbalanced data streams, but are relatively less robust; in contrast, ensemble learning approaches are typically more robust, but may require more computational resources and time.

6.2 Concept drift

Concept drift in an imbalanced data stream refers to the phenomenon where the underlying data distribution, particularly the class distribution, changes over time. This poses a significant challenge for classifiers trained on imbalanced data as they may struggle to adapt to the evolving patterns and imbalanced ratios (Agrahari and Singh 2022; Lu et al. 2018).

Recently an increasing number of researchers have focused on this aspect and have proposed new research methods. For example, Jiao et al. (2022) addresses the challenges of concept drift and class imbalance in data streams by proposing a dynamic ensemble selection approach. It incorporates a novel synthetic minority oversampling technique (AnnSMOTE) to generate new minority instances, adapts base classifiers to changing concepts, and constructs an optimal combination of classifiers based on local performance. Korycki and Krawczyk (2021) introduced a taxonomy of obstacles in multi-class imbalanced data streams impacted by concept drift. They also put forth a trainable concept drift detector based on Restricted Boltzmann Machine, capable of independently monitoring multiple classes and detecting changes through reconstruction error. Liu et al. (2021) propose CALMID, an integrated active learning method for multiclass imbalanced stream data with concept drift, which combines an integrated classifier, a drift detector, a label sliding window, a sample sliding window and an initialised training sample sequence. CALMID addresses the challenges of multiclass imbalance and concept drift using a variable threshold uncertainty strategy and a novel sample weight formulation. Ren et al. (2018) proposed Gradual Resampling Ensemble (GRE). GRE selectively resamples previous minority examples using a DBSCAN clustering approach, avoiding influences from small disjuncts and outliers, and ensures that only minority examples with low probability of overlapping with the current majority set are selected.

For imbalanced data stream scenarios with conceptual drift, classifiers constructed based on sampling methods are likely to be overfitted, leading to inefficient drift adaptation. Dynamic ensemble selection can generate a variety of few instances based on the current distribution of data streams for providing more valuable information for classifying conceptual drift. However, this also increases the complexity of the model, requiring more runtime in the data space as well as in classifier design and selection.

6.3 Incremental learning

Incremental learning to deal with imbalanced data streams is the process of continually updating and adapting the learning model to deal with concept drift, a phenomenon where the underlying data distribution changes, while also addressing class imbalances in the data stream. Ditzler and Polikar (2012).

In response to the challenges posed by imbalanced data streams, researchers have increasingly turned to incremental learning techniques as effective solutions. By gradually incorporating new information, incremental learning algorithms can dynamically adjust their decision boundaries, assign appropriate weights to different samples or classes, and prioritize learning from the minority class. These adaptive mechanisms contribute to improving the overall performance and accuracy of the model when dealing with imbalanced data

streams. For example, Li et al. (2020) proposed Dynamic Updated Ensemble (DUE) algorithm addresses the challenges of concept drift and class imbalance in learning nonstationary data streams by incrementally updating the model one chunk at a time, prioritizing misclassified examples, adapting to different concept drifts, handling the switch from majority to minority class, and maintaining efficiency with a limited number of classifiers. Lu et al. (2017) proposed Dynamic Weighted Majority Imbalanced Learning (DWMIL) to address the challenges of concept drift and class imbalance in data streams. DWMIL is a block-based incremental learning approach that utilises an ensemble framework with dynamically weighted base classifiers, allowing stability in non-drifting streams and rapid adaptation to new concepts. It is fully incremental, does not require storage of previous data, uses a limited number of classifiers to maintain high efficiency and has a simple implementation using only one threshold parameter. Li et al. (2020) proposed a block-based dynamic update integration (DUE) that aims to highlight examples of misclassification during model updates by learning one block at a time without accessing previous data and adapting to multiple types of conceptual drift in a timely manner. DUE overcomes the limitations of existing techniques and addresses the challenges of conceptual drift and class imbalance in non-smooth data streams.

Incremental learning aims to incrementally update models to accommodate new instances, adjust decision boundaries, and mitigate the impact of concept drift and class imbalance on classification performance. The research area focuses on developing methods that can learn from streaming data in an online manner, make timely and accurate predictions, and adapt to changing data characteristics. The incremental learning approach allows for adaptive updating of the model with new chunks of data without re-training and tuning, avoiding cumbersome training iteration sessions and therefore saving a significant amount of runtime.

7 Evaluation indicators for imbalanced learning

In this section, we describe various types of evaluation metrics regarding imbalanced learning. Accuracy cannot be chosen as an evaluation index in imbalanced learning. The main reasons are as follows:(1) In an imbalanced dataset, if a model tends to predict the majority of classes, a relatively high accuracy rate can be obtained even if the minority of categories are completely ignored. This makes the accuracy insensitive to the categorization performance of the minority categories and leads to bias in the evaluation. (2) Even if a model's classification performance for a minority of categories is poor, as long as the classification accuracy for the majority of categories is high, the accuracy may still be high, leading to failure to recognize imbalances. The evaluation indicators relating to imbalanced learning are shown in Table 6.

In Table 6 most of the evaluation indicators are calculated on the basis of the confusion matrix (CM). CM is a table for evaluating the performance of a classification model, which includes metrics such as True Positive (TP), False Positive (FP), False Negative (FN), and True Negative (TN), which are used to analyze the classification accuracy and misclassification of the model. The main reason for this is that metrics such as AUC and G-mean are unaffected by imbalances in the class distribution, as they are calculated based on the entire ROC curve or different parts of the confusion matrix. This makes them suitable for dealing with situations where there are large differences in the number of positive and negative class samples. In the formulas for AUC and F1-score, Precision denotes accuracy. TP_r is defined as the true positive rate and TN_r indicates the true negative rate.

In the case of multi-category imbalanced learning, MAUC combines the AUC values of multiple classes to provide a global multi-class performance metric, which enables a more comprehensive evaluation of the model's performance on different classes. On the other

Table 6 Evaluation metrics for imbalanced classifications

Evaluation metrics	Formulas
Recall	$\frac{TP}{TP+FN}$ (1)
Specificity	$\frac{TN}{TN+FP}$ (2)
Sensitivity	$\frac{TP}{TN+FP}$ (3)
AUC	$\frac{(1+TP_r-FP_r)}{2}$ (4)
G-mean	$\sqrt{Recall \times Specificity}$ (5)
F1-score	$\frac{2*Recall*Precision}{Recall+Precision}$ (6)
MAUC	$\frac{2}{m(m-1)}\sum_{i<j}\frac{[A(i,j)+A(j,i)]}{2}$ (7)
G-mean for multiclass	$\left(\prod_{k=1}^{n}\text{Recall}_k\right)^{\frac{1}{n}}$ (8)

hand, for multiclass imbalance problems, the evaluation metric of G-mean tends to follow the expansion, mainly because it provides a unified metric without the need to consider each class individually. In Table 6, m denotes the total number of classes, $A(i, j)$ is the AUC between two classes computed from column i of matrix M. M denotes an $N \times m$ matrix.

8 Application scenarios of imbalanced learning

Numerous real-world applications of imbalanced learning have led to the development of methods for learning from imbalanced data. This section concentrates on recent research pertaining to the practical applications of imbalanced learning. Figure 7 illustrates the categorization of the primary application domains into seven major areas, each encompassing specific real-life application challenges.

8.1 Biomedical area

In the biomedical field, where imbalanced data distributions are typical and accurate identification of minority class instances is essential for accurate decision-making and patient care, imbalanced learning techniques have been successfully used for tasks like rare disease detection, cancer diagnosis, DNA identification, and others. Figure 8 illustrates the architecture of the application of imbalanced learning in the biological area.

8.1.1 Protein subcellular assays

Subcellular localization of human proteins is critical for understanding their function, diagnostic and prognostic studies of pathological conditions, and clinical decision-making, but multi-label classifiers are challenged by severe bias and reduced predictive power when dealing with proteins that are present in multiple locations simultaneously due to data imbalance

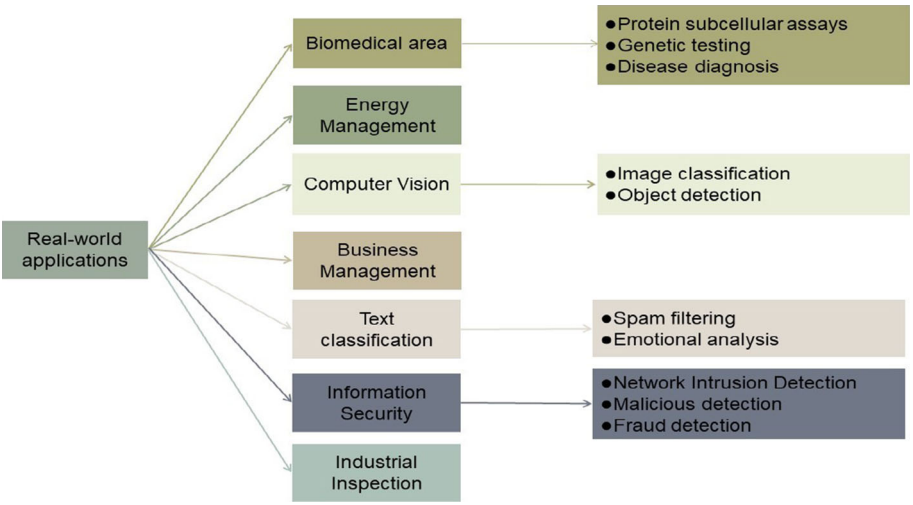

Fig. 7 Real-world Application classification

(Rana et al. 2023). Wang and Wei (2022) used imbalanced multilabeling of immunohistochemical images to make protein subcellular localization predictions. Rana et al. (Ahsan et al. 2022) use a multi-label oversampling approach to cope with this type of problem. Protein classification includes the latest solutions recently proposed by Yin et al. (2022) and Hung et al. (2022).

8.1.2 Genetic testing

Genetic testing now makes use of imbalanced learning techniques to handle class imbalance problems brought on by the unequal distribution of genetic variations or profiles, hence

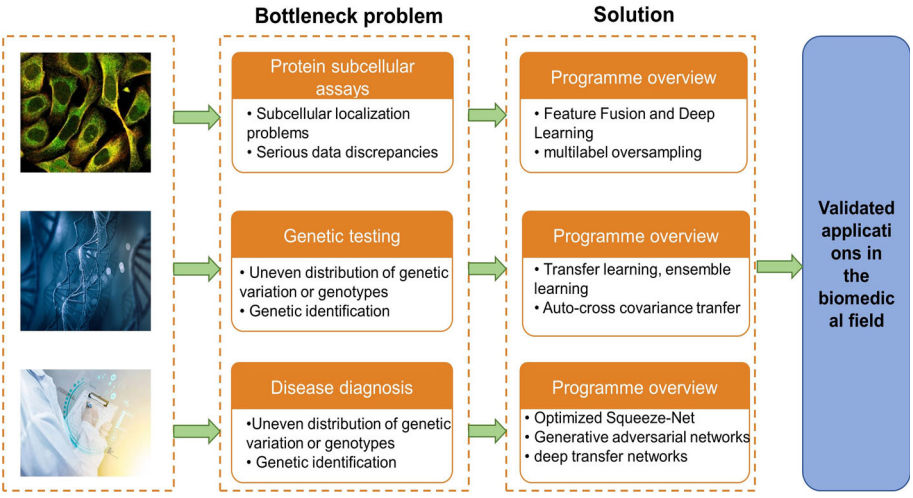

Fig. 8 Architecture of possible applications in the biological area

enhancing the accuracy and dependability of genetic analysis and identification procedures. Wang et al. (2015) solved the classification of miRNA imbalance data using an ensemble learning approach. Li et al. (2023) and Liu et al. (2016) combine integrated learning methods with other models to address DNA-binding protein identification.

8.1.3 Disease diagnosis

Early diagnosis of cancer is crucial for patients, yet data imbalance and quality imbalance between majority and minority classes leading to misclassification is a huge challenge in medical data analysis. Although majority class samples and correct classification are more important for classifiers, cancer diagnosis relies on minority class samples. Researchers are therefore conducting a comprehensive study of the imbalanced data problem from a medical perspective in order to explore new approaches to cancer diagnosis (Fotouhi et al. 2019). Fotouhi et al. (2019) and Behrad and Abadeh (2022) provide a summary of machine learning and deep learning approaches in tackling disease diagnosis imbalances and present future challenges. Various studies (Xiao et al. 2021; Saini and Susan 2022; Singh et al. 2020; Saini and Susan 2020) have proposed novel approaches for imbalanced data in disease diagnosis, including the use of optimized SqueezeNet with bald eagle search optimization, generative adversarial networks, deep transfer networks, and transfer learning with minority data augmentation, all showing promising results in improving classification accuracy for imbalanced breast cancer and melanoma datasets.

8.2 Information Security

Imbalanced learning techniques are finding valuable applications in the field of information security. With the increasing complexity and sophistication of cyber threats, traditional classification models often struggle to effectively handle imbalanced datasets in tasks such as network intrusion detection, malicious detection and fraud detection. Imbalanced learning techniques can identify rare and critical security events, enhance the accuracy of anomaly detection, and help detect previously invisible threats. Figure 9 illustrates the architecture of the application of imbalanced learning in the information Security area.

8.2.1 Network intrusion detection

Imbalanced learning has shown significant potential in the field of Network Intrusion Detection (NID). NID aims to identify and prevent unauthorized access or malicious activities within a computer network. However, traditional detection methods often struggle to handle imbalanced datasets, where the majority of network traffic is normal while a small portion represents intrusions. Imbalanced learning techniques provide effective solutions by addressing the class imbalance problem and improving the detection performance on minority class instances. A number of research scholars have explored the landscape of intrusion detection techniques in recent years, providing insights into application domains, attack detection techniques, evaluation metrics and datasets (Yang et al. 2022; Di Mauro et al. 2021). It also discusses the concept of intrusion detection systems (IDS), presents a classification of machine learning (ML) and deep learning (DL) techniques used in network-based IDS systems, and highlights the advantages, limitations of ML and DL-based IDSs. Cui et al. (2023) proposed a new multi-module integrated intrusion detection system. Recently, some scholars have proposed dynamic integration-based methods, resampling-based methods and

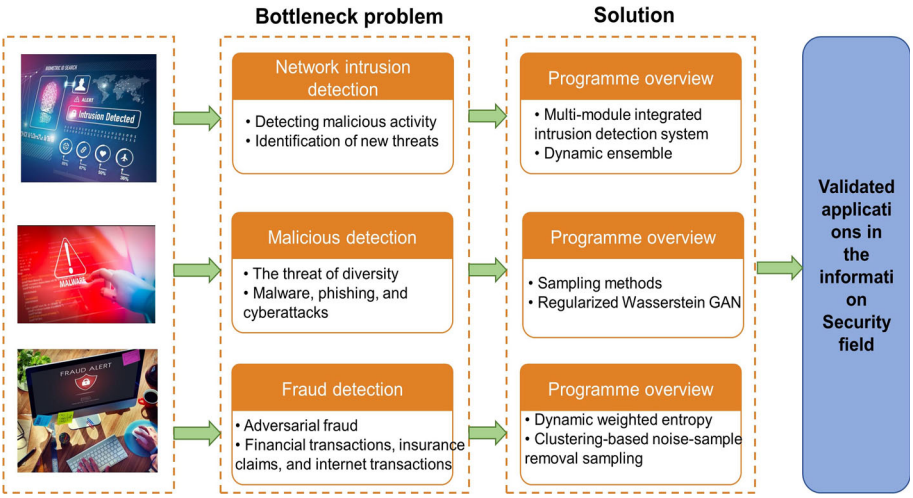

Fig. 9 Architecture of the application in the information security area

combining LSTM with other techniques to deal with the class imbalance problem in network intrusion detection (Ren et al. 2023; Bagui and Li 2021; Gupta et al. 2021).

8.2.2 Malicious detection

The goal of malicious detection is to recognize and counteract different forms of criminal activity, such as malware, phishing, and cyberattacks. Imbalanced learning is a critical component of this process. By leveraging imbalanced learning, malicious detection systems can achieve better performance in identifying and mitigating malicious activities, enhancing the security of computer systems, networks, and online platforms. Identification of malicious domains is of great importance as they are one of the main resources necessary for conducting cyber attacks. Zhauniarovich et al. (2018); Sharma and Rattan (2021) systematically studied and synthesized the existing methods differing in terms of data sources, analysis methods and evaluation strategies. Phung and Mimura (2021) propose a new approach to detect malicious JavaScript using machine learning techniques and oversampling methods. Chapaneri and Shah (2022) proposed regularized Wasserstein generative adversarial networks (WGAN) to solve the imbalanced malicious detection problem. Cui et al. (2021) proposed a multi-objective Restricted Boltzmann Machine (RBM) model combined with Non-Dominated Sorting Genetic Algorithm (NSGA-II) to solve the malicious code attack problem, and also proposed to find an efficient malicious code detection method.

8.2.3 Fraud detection

In the field of fraud detection, where it is important to spot fraudulent activity in a variety of contexts, including financial transactions, insurance claims, and internet transactions, imbalanced learning plays a key role. Imbalanced learning techniques provide an effective solution by addressing the category imbalance problem and improving the detection performance for a small number of fraudulent instances. The literature (Pourhabibi et al. 2020) surveys current trends and identifies key challenges that require significant research efforts to improve the trustworthiness of the technique, while review and statistical machine learning techniques

have a wide range of applications in fraud detection as e-commerce systems increase and financial transactions become online with increasing fraud, prevention techniques are effective but fraud detection methods are critical. Li et al. (2021) proposed a hybrid approach to imbalance fraud detection with dynamic weighted entropy. Zhu et al. (2023) proposed a clustering-based noise-sample removal undersampling scheme (NUS) for imbalanced credit card fraud detection.

8.3 Computer Vision

The discipline of computer vision has made substantial use of imbalanced learning, notably for tasks like object detection and image classification. By leveraging imbalanced learning in computer vision, researchers and practitioners can improve the robustness and reliability of vision-based systems, enabling applications in areas such as surveillance, medical imaging, autonomous vehicles, and object recognition in diverse real-world scenarios. Figure 10 illustrates the architecture of the application of imbalanced learning in the computer vision area.

8.3.1 Image classification

In image classification tasks, it is common to have imbalanced distributions of images across different classes, where some classes may have significantly fewer samples than others. This imbalance can negatively impact the performance of conventional classification models, leading to biased predictions and lower accuracy on minority classes. Over the past few years, more and more scholars have created new imbalanced learning algorithms for solving imbalanced image classification. For example, Wang et al. (2021) introduced Deep Attention-based Imbalanced Image Classification (DAIIC) to automatically allocate more attention to few classes in a data-driven manner. Huang et al. (2019) mitigate the face classification problem for class-imbalanced data by classical strategies (such as class resampling or cost-sensitive training) and by enforcing deep networks to learn more discriminative deep representations. Jin et al. (2022) proposed the Balanced Active Learning (BAL) method to alleviate class imbalance by compensating for a few classes of labeled queries, achieving state-of-the-art active learning performance on an imbalanced image classification dataset.

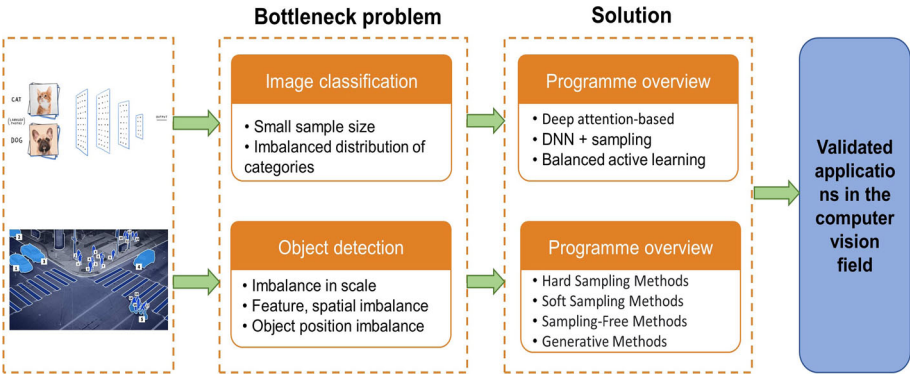

Fig. 10 Architecture of the application in the computer vision

8.3.2 Object detection

Object detection involves the identification and localization of objects within images or videos. However, in many real-world scenarios, the distribution of objects across different classes can be highly imbalanced, where some classes have significantly fewer instances than others. This class imbalance poses a challenge for object detection models as they may struggle to accurately detect and classify objects from the minority classes. Oksuz et al. (2020) provide a systematic review of imbalance problems in target detection, introduce a problem-based taxonomy, discuss each problem in detail, and provide a critical view of existing solutions. Huang and Liu (2022) proposed a dense detector for small target detection, which solves the scale imbalance of samples and features by Libra ellipse sampling and residual low-level feature enhancement.

8.4 Business management

In business management, imbalanced learning is widely used for user churn prediction and consumer behavior analysis. Churn prediction refers to predicting which users are likely to stop using an organization's products or services, which is important for organizations because they can take steps to retain these users (Haixiang et al. 2017). Consumer behavior analytics aims to understand and predict consumers' purchasing decisions and preferences to help companies develop effective marketing strategies. However, since the ratio of purchasers to non-purchasers is usually unbalanced, imbalanced learning can help solve this problem by improving the ability to identify and predict purchasers, thus optimizing marketing campaigns and improving sales. For example, using customer data to identify prospects who are more likely to buy caravan insurance (Almas et al. 2012). Literature (Wu and Meng 2016; De Caigny et al. 2018) using various consumer data to analyze customer behavioral characteristics and habituation.

8.5 Text classification

Imbalanced learning is widely used in spam filtering and sentiment analysis. In spam filtering, traditional classification algorithms may be ineffective in recognizing spam due to the relatively small number of spam emails and the large number of normal emails. In sentiment analysis, for which the distribution of samples representing different sentiment polarities (such as positive, negative and neutral) in textual data is usually uneven, imbalanced learning can be used to address this problem and improve the ability to accurately predict and recognize the sentiments of a few categories. Figure 11 illustrates the architecture of the application of imbalanced learning in the text classification area.

8.5.1 Spam filtering

Spam filtering is essentially an unbalanced classification task designed to identify and block emails that are useless or fraudulent in nature in order to protect users from the nuisance and threat of spam. Francisco et al. (Jáñez-Martino et al. 2023) highlighted the challenges in developing robust spam email filters, including the dynamic nature of the environment and the presence of spammers as adversarial figures, and provides an analysis of spammer strategies and state-of-the-art machine learning techniques. Barushka and Hajek (2018) proposed

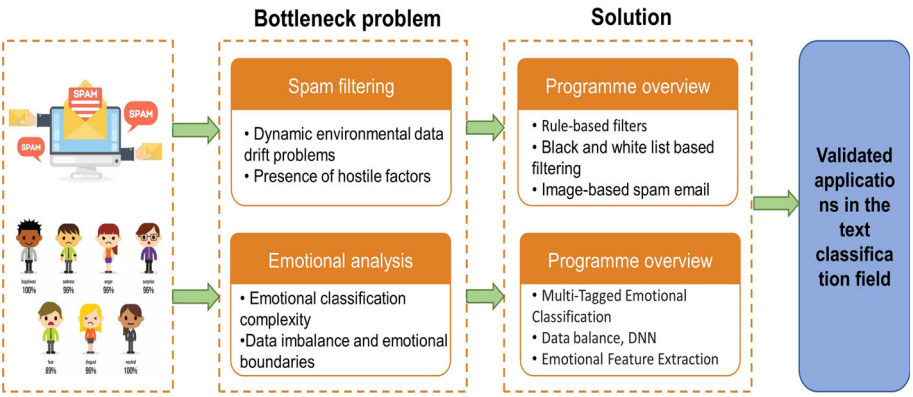

Fig. 11 Architecture of the application in the text classification

a regularised deep multilayer perceptron NN model (DBB-RDNN-ReL) based on feature selection and rectified linear units for spam filtering. Rao et al. (2023) suggested a hybrid framework for identifying social media spam that incorporates dataset balance, sophisticated word embedding techniques, machine learning, and deep learning methodologies, as well as the self-attention mechanism.

8.5.2 Emotional analysis

Sentiment analysis aims to identify and understand the emotional tendencies expressed in a text to determine whether they are positive, negative or neutral. Deng and Ren (2023) provide a comprehensive overview of recent advances in deep neural network text emotion recognition (TER), covering all aspects of word embedding, architecture and training. It highlights remaining challenges and opportunities in the areas of dataset availability, sentiment boundaries, extractable sentiment information, and TER in conversations. Ding et al. (2021) proposed a feature extraction-based EEG emotion recognition method using dispersion entropy of different frequency bands of EEG signals and data balancing using a random oversampling algorithm, which resulted in improved characterisation and faster recognition compared to other methods. Lin et al. (2022a, b) proposed novel multi-label sentiment categorization that experimentally validates and solves the class imbalance problem.

8.6 Energy management

The application of imbalanced learning in energy management, especially in the area of electricity theft detection, can help to solve the imbalance problem of electricity theft events in energy systems. The researchers aim to study and analyze energy consumption data and related characteristics to identify and monitor electricity theft by a small number of users. For example, Yan and Wen (2021) proposed a power theft detector using Extreme Gradient Boosting (XGBoost) based metering data. Cai et al. (2023) proposed an integrated learning model based on random forest and weighted support vector description to analyze the problem of electricity theft detection in a complex grid environment. Pereira and Saraiva (2021)explored the application of Convolutional Neural Networks (CNNs) combined with various techniques for balancing imbalanced datasets to detect power theft activity.

8.7 Industrial Inspection

The application of imbalanced learning in industrial inspection is mainly focused on the field of anomaly detection and fault diagnosis (Yang et al. 2023, 2022; Chen et al. 2021). The problem of class imbalance results from the fact that in industrial environments, normal samples usually dominate, while abnormal or faulty samples are relatively few. Imbalanced learning can effectively enable the model to better learn and recognize anomalous or faulty samples during the training process. Ren et al. (2023) and Zhang et al. (2022) offered a thorough evaluation of research accomplishments in the field of fault detection under data imbalance, including data processing methods, model creation methods, and training optimization approaches. Kuang et al. (2021) proposed a class-imbalanced adversarial transfer learning (CIATL) network to address the challenge of cross-domain troubleshooting in the presence of limited availability of class-balanced data. Liu et al. (2021) proposed a novel unbalanced data classification method based on weakly supervised learning, which utilizes the Bagging algorithm to generate balanced subsets and employs Support Vector Machine (SVM) classification as a way to improve fault diagnosis performance. Recently there are also some meta-autonomous learning methods based on multi-view sampling, deep reinforcement learning algorithms to optimize the sample distribution and dynamically balanced domain adversarial network algorithms to solve the anomaly detection problem (Lyu et al. 2022; Fan et al. 2021; Ren et al. 2022).

9 Future research directions

Based on a comprehensive review of relevant algorithms and an analysis of their strengths, weaknesses, runtimes, and model complexity discussed in the previous sections, this section will now explore the current challenges and future perspectives of imbalanced learning. These new research directions provide perspectives for a deeper understanding of the complex issues in imbalanced learning, help to address the challenges that have emerged in recent years, and promote more comprehensive and in-depth progress in imbalanced learning research. Building upon the analyses presented earlier, we aim to provide a deeper understanding of the topic at hand.

Increased adaptability in transfer learning scenarios Transfer learning focuses on exploring feature representation discriminability, but they deal with domain alignment and class discriminability independently. Striking a balance between alignment and discriminability is critical, as overemphasising one can lead to the loss of the other. If the property of sample size imbalance between domains is ignored, this can lead to bias, poor alignment and discrimination during training, and ultimately to negative migration (Li et al. 2023). As we discussed earlier, the nature of the data distribution over time can affect the performance of migration learning.

Current approaches (Singh et al. 2020; Liu et al. 2023) predominantly address adaptations to conceptual drift and static category imbalances. However, it is worth exploring the feasibility of predicting changes in advance. Can we anticipate the evolution of category imbalances over time and develop proactive methods to effectively respond? Such proactive measures would substantially reduce the recovery time following shifts in imbalanced data streams and yield more resilient classifiers.

Active learning for imbalanced data Active learning faces unique challenges when dealing with imbalanced data. Suppose there is a prediction problem formulation: $p(\bar{y}_n, \tilde{y}_n|x) =$

$\frac{p(y_n|\tilde{y}_n,x)p(x|\tilde{y}_n)p(\tilde{y}_n)}{\sum_{n=1}^{Nc} p(x|\tilde{y}_n)p(\tilde{y}_n)}$, $p(\tilde{y}_n)$ is the probability that the model predicts a sample to be in a certain class, $p(x|\tilde{y}_n)$ is the corresponding generative model. Both of the above can be easily calculated. But $p(y_n|\tilde{y}_n, x)$ is sample x. It is impractical to calculate the probability of an instance being truly the n^{th} class when the model predicts it to be the n^{th} class, as the true label remains unknown in such cases. Active learning algorithms universally require a model trained initially on a randomly chosen sample, serving as a basis for discussing uncertainty or feature space coverage. However, when dealing with minority classes, a random selection often results in a minimal or even zero inclusion of samples. Employing such a subset for training leads to the model producing inaccurate yet highly confident predictions for these infrequent samples, causing uncertainty-based methods to lean towards avoiding their selection (Liu et al. 2021; Aguiar and Cano 2023). Regarding feature space coverage, the limited number of minority class samples, ineffective feature separation from the main classes post-training, and the mingling of majority class samples make feature space coverage-based methods more inclined to choose majority class samples to cover that particular area.

Moreover, the scarcity of minority class samples can introduce significant noise during the labeling process, thereby impacting the classifier's performance. Additionally, the class imbalance may introduce bias in the active learning process, causing the classifier to prioritize majority class samples while overlooking the significance of minority class samples (Aguiar and Cano 2023). To address these challenges, active learning strategies and algorithms specifically tailored for imbalanced data are required. These approaches aim to enhance classifier performance and maximize the utilization of valuable information from minority class.

Addressing class imbalance in federated learning Federated learning (FL) utilises heterogeneous edge devices and a central server for collaborative learning, where local model training is performed by keeping the collected data locally rather than transmitting the data directly. The edge devices transmit the trained models to a central server, which then performs global model aggregation. Although FL performs well in handling cross-device data migration, its performance is typically poorer when training imbalanced data compared to standard centrally learnt models (Duan et al. 2019). The complexity of data imbalances, which can occur locally on a device or across multiple devices, challenges federated learning techniques, particularly the need to balance the need to address data imbalances with the need to maintain data privacy. This is at the same time that the imbalanced data flow problem mentioned in Section 6 also affects FL performance. Duan et al. (2019) have shown mathematically and theoretically that the data imbalance property can seriously affect the performance of federated learning.

The main reasons why FL is challenging when encountering imbalanced data are as follows:(1) Wang et al. (2021) have previously pointed out that FL, due to its decentralised nature, may suffer from class imbalances at different levels, including at the local level as well as at the global level across one or more client devices of multiple clients, constituting a system-level imbalance problem. (2) Mismatch in class distribution between clients and servers may also trigger degradation in FL performance. Differences in class distribution among all clients participating in FL can degrade the overall performance of the global model and increase convergence latency. Addressing these issues requires effective management and tuning of class imbalances at different levels and distribution mismatches between clients and servers to improve the robustness and performance of the FL system.

Addressing dynamic class changes in data streams Aguiar et al. (2023) have demonstrated experimentally that many current learning methods for coping with static unbalanced

data perform poorly in unbalanced data streams, especially when confronted with multiple categories. While numerous researchers have examined the impact of class quantity in imbalance problems, the issue of dynamic changes in the number of classes remains largely unexplored. In real-world scenarios, classes may emerge, vanish, and reoccur over time. The dynamic nature of class changes, coupled with the presence of class imbalance, presents a formidable challenge that necessitates the development of flexible models. Such models should possess the ability to detect new classes and seamlessly incorporate them into the model structure, as well as forget obsolete classes while retaining knowledge of recurring classes.

The main problem is the natural data stream shift, which changes as the relevance of the data that generates the minority stream changes, leading to performance degradation. It may be possible in the future to develop programmes to detect and measure dataset shifts, but the real challenge is how to tune it to focus more on the minority class. Addressing this issue requires a deeper understanding of the changing patterns of data streams and the development of methods that can be effectively tuned to maintain focus on the minority class. This is critical to ensure that performance is not negatively impacted by data flow shifts.

Addressing the multi-modal imbalanced learning problem Multimodal learning achieves comprehensive perception and understanding by understanding different types of data. Most of the currently available modal learning methods usually give the same importance to the features of each modality, and multimodal models tend to make use of modal data with smaller values of the loss function for parameter updating during the optimisation process. This causes the model to be biased towards one of the dominant modalities during algorithm training (Behrad and Abadeh 2022). This is mainly due to the fact that the modal data with better performance inhibits the role of the other weakly performing modal data in the update, thus creating an imbalance problem.

To date, few studies have simultaneously addressed the challenges associated with class imbalanced and multimodal data (Sleeman et al. 2022). When confronted with multimodal data, models must prioritize relevant features extracted from each data domain (Kim and Sohn 2020). This emphasis on capturing pertinent information from multiple modalities can introduce sensitivities, particularly when working with a limited number of classes.

Addressing the multi-label cross-domain imbalanced learning problem In multi-label learning (MLD), the imbalance problem covers three levels: intra-label, inter-label and label set. Intra-label imbalance refers to fewer positive samples in individual labels, inter-label imbalance involves differences in the number of positive classes in independent labels, and label set imbalance is affected by label sparsity, which leads to more frequent occurrence of certain label sets. There has been relatively limited research on the problem of imbalanced multi-label classification, with existing work on resolving the imbalance focusing on the area of single-label classification (Rana et al. 2023; Tarekegn et al. 2021). Despite the increasing demand for multi-label classification in different domains, a comprehensive framework to effectively deal with the imbalance problem in multi-label classification has not yet been fully investigated.

More-ever, In the context of multi-label multimodal datasets, the presence of class correlations and modal variances poses a significant challenge (Tarekegn et al. 2021). Such datasets exhibit inter-label correlations and intra-modality differences, adding complexity to model learning and prediction tasks. Therefore, it becomes crucial to accurately capture label correlations and modal differences to enhance categorization and prediction accuracy.

10 Conclusion

In this survey, we present a comprehensive analysis of the challenges posed by class imbalance. We examine its characteristics and associated issues, and propose a novel classification approach to comprehensively review and analyze the existing methods for addressing imbalanced learning. Notably, unlike previous surveys in the field, which primarily focus on data mining and machine learning perspectives, we also delve into the advancements of imbalanced learning in the context of deep learning, specifically long-tail learning. Furthermore, we explore the current state of research on the practical application of imbalanced learning in seven distinct domains, and identify new research directions and areas of innovation for methods and tasks. Our aim is to provide researchers and the community with a comprehensive understanding of imbalanced learning, thereby facilitating future research endeavors in this domain.

Acknowledgements This work was supported in part by National Key R&D Pogram of China 2023YFA1011601, and in part by the Major Key Project of PCL, China under Grant PCL2023AS7-1, and in part by the National Natural Science Foundation of China No. U21A20478, 62106224, U21B2029, and in part by the Open Research Project KFKT2022B11 of the State Key Lab. for Novel Software Technology, Nanjing University, China.

Author Contributions Wuxing Chen (WC) and Kaixiang Yang (KY) conceived the research idea and designed the study. WC, KY, Zhiwen Yu (ZY), Yifan Shi (YS), and C. L. Philip Chen (PC) analyzed the research. WC and KY wrote the main manuscript text. ZY and YS prepared figures 1-6. All authors reviewed and provided critical feedback on the manuscript. All authors contributed to discussions regarding the research and revised the manuscript for intellectual content.

Declarations

Competing interests The authors declare no competing interests.

References

Abedin MZ, Guotai C, Hajek P, Zhang T (2022) Combining weighted smote with ensemble learning for the class-imbalanced prediction of small business credit risk. Complex Intell Syst, 1–21

Agrahari S, Singh AK (2022) Concept drift detection in data stream mining: a literature review. Journal of King Saud University-Computer and Information Sciences 34(10):9523–9540

Aguiar G, Cano A (2023) An active learning budget-based oversampling approach for partially labeled multi-class imbalanced data streams. In: Proceedings of the 38th ACM/SIGAPP symposium on applied computing, pp 382–389

Aguiar G, Krawczyk B, Cano A (2023) A survey on learning from imbalanced data streams: taxonomy, challenges, empirical study, and reproducible experimental framework. Mach Learn, 1–79

Ahsan R, Ebrahimi F, Ebrahimi M (2022) Classification of imbalanced protein sequences with deep-learning approaches; application on influenza a imbalanced virus classes. Inform Med Unlocked 29:100860

Akbani R, Kwek S, Japkowicz N (2004) Applying support vector machines to imbalanced datasets. In: Machine learning: ECML 2004: 15th European conference on machine learning, Pisa, Italy, September 20-24, 2004. Proceedings 15. Springer, pp 39–50
Akila S, Reddy US (2018) Cost-sensitive risk induced bayesian inference bagging (ribib) for credit card fraud detection. J Comput Sci 27:247–254
Alfhaid MA, Abdullah M (2021) Classification of imbalanced data stream: techniques and challenges. Artif Intell 9(2):36–52
Almas A, Farquad M, Avala NR, Sultana J (2012) Enhancing the performance of decision tree: a research study of dealing with unbalanced data. In: Seventh international conference on digital information management (ICDIM 2012). IEEE, pp 7–10
Bader-El-Den M, Teitei E, Perry T (2018) Biased random forest for dealing with the class imbalance problem. IEEE Trans Neural Netw Learn Syst 30(7):2163–2172
Bagui S, Li K (2021) Resampling imbalanced data for network intrusion detection datasets. J Big Data 8(1):1–41
Barushka A, Hajek P (2018) Spam filtering using integrated distribution-based balancing approach and regularized deep neural networks. Appl Intell 48:3538–3556
Behrad F, Abadeh MS (2022) An overview of deep learning methods for multimodal medical data mining. Expert Syst Appl 200:117006
Błaszczyński J, Stefanowski J (2015) Neighbourhood sampling in bagging for imbalanced data. Neurocomputing 150:529–542
Branco P, Torgo L, Ribeiro RP (2019) Preprocessing approaches for imbalanced distributions in regression. Neurocomputing 343:76–99
Branco P, Torgo L, Ribeiro RP (2017) Smogn: a pre-processing approach for imbalanced regression. In: First international workshop on learning with imbalanced domains: theory and applications. PMLR, pp 36–50
Branco P, Torgo L, Ribeiro RP (2018) Rebagg: resampled bagging for imbalanced regression. In: Second international workshop on learning with imbalanced domains: theory and applications. PMLR, pp 67–81
Breiman L (1996) Bagging predictors. Mach Learn 24:123–140
Bunkhumpornpat C, Sinapiromsaran K, Lursinsap C (2009) Safe-level-smote: safe-levelsynthetic minority over-sampling technique for handling the class imbalanced problem. In: Advances in knowledge discovery and data mining: 13th Pacific-Asia Conference, PAKDD 2009 Bangkok, Thailand, April 27-30, 2009 Proceedings 13. Springer, pp 475–482
Cai L, Wang H, Jiang F, Zhang Y, Peng Y (2022) A new clustering mining algorithm for multi-source imbalanced location data. Inf Sci 584:50–64
Cai Q, Li P, Wang R (2023) Electricity theft detection based on hybrid random forest and weighted support vector data description. Int J Electr Power Energy Syst 153:109283
Cano A, Krawczyk B (2022) Rose: robust online self-adjusting ensemble for continual learning on imbalanced drifting data streams. Mach Learn 111(7):2561–2599
Cao B, Liu Y, Hou C, Fan J, Zheng B, Yin J (2020) Expediting the accuracy-improving process of svms for class imbalance learning. IEEE Trans Knowl Data Eng 33(11):3550–3567
Cao B, Liu Y, Hou C, Fan J, Zheng B, Yin J (2021) Expediting the accuracy-improving process of svms for class imbalance learning. IEEE Trans Knowl Data Eng 33(11):3550–3567
Cao K, Wei C, Gaidon A, Arechiga N, Ma T (2019) Learning imbalanced datasets with label-distribution-aware margin loss. In: Proceedings of the 33rd international conference on neural information processing systems, pp 1567–1578
Castro CL, Braga AP (2013) Novel costsensitive approach to improve the multilayer perceptron performance on imbalanced data. IEEE Trans Neural Netw Learn Syst 24(6):888–899
Chapaneri R, Shah S (2022) Enhanced detection of imbalanced malicious network traffic with regularized generative adversarial networks. J Netw Comput Appl 202:103368
Chawla NV, Lazarevic A, Hall LO, Bowyer KW (2003) Smoteboost: improving prediction of the minority class in boosting. In: Knowledge Discovery in Databases: PKDD 2003: 7th European conference on principles and practice of knowledge discovery in databases, Cavtat-Dubrovnik, Croatia, September 22-26, 2003. Proceedings 7. Springer, pp 107–119
Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP (2002) Smote: synthetic minority over-sampling technique. J Artif Intell Res 16:321–357
Chen J, Xiu Z, Goldstein B, Henao R, Carin L, Tao C (2021) Supercharging imbalanced data learning with energy-based contrastive representation transfer. Adv Neural Inf Process Syst 34:21229–21243
Chen W, Yang K, Yu Z, Zhang W (2022a) Double-kernel based class-specific broad learning system for multiclass imbalance learning. Knowl-Based Syst 253:109535
Chen W, Yang K, Zhang W, Shi Y, Yu Z (2022b) Double-kernelized weighted broad learning system for imbalanced data. Neural Comput Appl 34(22):19923–19936

Chen W, Yang K, Shi Y, Feng Q, Zhang C, Yu Z (2021) Kernel-based classspecific broad learning system for software defect prediction. In: 2021 8th International conference on information, cybernetics, and computational social systems (ICCSS). IEEE, pp 109–114
Chen X, Zhou Y, Wu D, Zhang W, Zhou Y, Li B, Wang W (2022) Imagine by reasoning: a reasoning-based implicit semantic data augmentation for long-tailed classification. In: Proceedings of the AAAI conference on artificial intelligence, vol 36, pp 356–364
Choudhary R, Shukla S (2021) A clustering based ensemble of weighted kernelized extreme learning machine for class imbalance learning. Expert Syst Appl 164:114041
Cui Z, Zhao Y, Cao Y, Cai X, Zhang W, Chen J (2021) Malicious code detection under 5g hetnets based on a multi-objective rbm model. IEEE Network 35(2):82–87
Cui J, Zong L, Xie J, Tang M (2023) A novel multi-module integrated intrusion detection system for high-dimensional imbalanced data. Appl Intell 53(1):272–288
Cui Y, Jia M, Lin T-Y, Song Y, Belongie S (2019) Class-balanced loss based on effective number of samples. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 9268–9277
Cui J, Zhong Z, Liu S, Yu B, Jia J (2021) Parametric contrastive learning. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 715–724
Dablain DA, Bellinger C, Krawczyk B, Chawla NV (2023) Efficient augmentation for imbalanced deep learning. In: 2023 IEEE 39th international conference on data engineering (ICDE). IEEE, pp 1433–1446
Dablain D, Krawczyk B, Chawla NV (2022) Deepsmote: fusing deep learning and smote for imbalanced data. IEEE Trans Neural Netw Learn Syst
Datta S, Das S (2015) Near-bayesian support vector machines for imbalanced data classification with equal or unequal misclassification costs. Neural Netw 70:39–52
Datta S, Das S (2019) Multiobjective support vector machines: handling class imbalance with pareto optimality. IEEE Trans Neural Netw Learn Syst 30(5):1602–1608
Datta S, Ghosh A, Sanyal K, Das S (2017) A radial boundary intersection aided interior point method for multi-objective optimization. Inf Sci 377:1–16
Datta S, Nag S, Das S (2020) Boosting with lexicographic programming: addressing class imbalance without cost tuning. IEEE Trans Knowl Data Eng 32(5):883–897
De Caigny A, Coussement K, De Bock KW (2018) A new hybrid classification algorithm for customer churn prediction based on logistic regression and decision trees. Eur J Oper Res 269(2):760–772
Deng J, Ren F (2023) A survey of textual emotion recognition and its challenges. IEEE Trans Affect Comput 14(1):49–67
Desai A, Wu T-Y, Tripathi S, Vasconcelos N (2021) Learning of visual relations: the devil is in the tails. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 15404–15413
Di Mauro M, Galatro G, Fortino G, Liotta A (2021) Supervised feature selection techniques in network intrusion detection: a critical review. Eng Appl Artif Intell 101:104216
Ding X-W, Liu Z-T, Li D-Y, He Y, Wu M (2021) Electroencephalogram emotion recognition based on dispersion entropy feature extraction using random oversampling imbalanced data processing. IEEE Trans Cogn Dev Syst 14(3):882–891
Ditzler G, Polikar R (2012) Incremental learning of concept drift from streaming imbalanced data. IEEE Trans Knowl Data Eng 25(10):2283–2301
Dixit A, Mani A (2023) Sampling technique for noisy and borderline examples problem in imbalanced classification. Appl Soft Comput 142:110361
Dong X, Yu Z, Cao W, Shi Y, Ma Q (2020) A survey on ensemble learning. Front Comp Sci 14:241–258
Douzas G, Bacao F (2017) Self-organizing map oversampling (somo) for imbalanced data set learning. Expert Syst Appl 82:40–52
Douzas G, Bacao F, Last F (2018) Improving imbalanced learning through a heuristic oversampling method based on k-means and smote. Inf Sci 465:1–20
Du H, Zhang Y, Gang K, Zhang L, Chen Y-C (2021) Online ensemble learning algorithm for imbalanced data stream. Appl Soft Comput 107:107378
Duan M, Liu D, Chen X, Tan Y, Ren J, Qiao L, Liang L (2019) Astraea: selfbalancing federated learning for improving classification accuracy of mobile deep learning applications. In: 2019 IEEE 37th International conference on computer design (ICCD). IEEE, pp 246–254
Fan Q, Wang Z, Li D, Gao D, Zha H (2017) Entropy-based fuzzy support vector machine for imbalanced datasets. Knowl-Based Syst 115:87–99
Fan S, Zhang X, Song Z (2021) Imbalanced sample selection with deep reinforcement learning for fault diagnosis. IEEE Trans Industr Inf 18(4):2518–2527
Fan W, Stolfo SJ, Zhang J, Chan PK (1999) Adacost: misclassification cost-sensitive boosting. In: Icml, vol 99, pp 97–105

Feng B, Gu Y, Yu H, Yang X, Gao S (2022) Dme: an adaptive and just-in-time weighted ensemble learning method for classifying block-based concept drift steam. IEEE Access 10:120578–120591
Fernández A, García S, Galar M, Prati RC, Krawczyk B, Herrera F, Fernández A, García S, Galar M, Prati RC et al (2018) Learning from imbalanced data streams. Learning from imbalanced data sets, 279–303
Fotouhi S, Asadi S, Kattan MW (2019) A comprehensive data level analysis for cancer diagnosis on imbalanced data. J Biomed Inform 90:103089
Freund Y, Schapire RE (1997) A decisiontheoretic generalization of on-line learning and an application to boosting. J Comput Syst Sci 55(1):119–139
Galar M, Fernandez A, Barrenechea E, Bustince H, Herrera F (2012) A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches. IEEE Trans Syst, Man, and Cybernetics, Part C (Applications and Reviews) 42(4):463–484
Ghosh K, Bellinger C, Corizzo R, Branco P, Krawczyk B, Japkowicz N (2022) The class imbalance problem in deep learning. Mach Learn, 1–57
Guo Y, Feng J, Jiao B, Cui N, Yang S, Yu Z (2022) A dual evolutionary bagging for class imbalance learning. Expert Syst Appl 206:117843
Gupta N, Jindal V (2021) Bedi P (2021) Lio-ids: handling class imbalance using lstm and improved one-vs-one technique in intrusion detection system. Comput Netw 192:108076
Gutiérrez-Tobal GC, Álvarez D, Vaquerizo-Villar F, Crespo A, Kheirandish-Gozal L, Gozal D, Campo F, Hornero R (2021) Ensemble-learning regression to estimate sleep apnea severity using at-home oximetry in adults. Appl Soft Comput 111:107827
Haixiang G, Yijing L, Shang J, Mingyun G, Yuanyue H, Bing G (2017) Learning from class-imbalanced data: review of methods and applications. Expert Syst Appl 73:220–239
Halder B, Hasan KA, Amagasa T, Ahmed MM (2023) Autonomic active learning strategy using cluster-based ensemble classifier for concept drifts in imbalanced data stream. Expert Syst Appl 120578
Han M, Guo H, Li J, Wang W (2023) Globallocal information based oversampling for multi-class imbalanced data. Int J Mach Learn Cybern 14(6):2071–2086
Han H, Wang W-Y, Mao B-H (2005) Borderline-smote: a new over-sampling method in imbalanced data sets learning. In: Advances in intelligent computing: international conference on intelligent computing, ICIC 2005, Hefei, China, August 23-26, 2005, Proceedings, Part I 1. Springer, pp 878–887
He H, Garcia EA (2009) Learning from imbalanced data. IEEE Trans Knowl Data Eng 21(9):1263–1284
He H, Bai Y, Garcia EA, Li S (2008) Adasyn: adaptive synthetic sampling approach for imbalanced learning. In: 2008 IEEE International joint conference on neural networks (IEEE World Congress on Computational Intelligence). IEEE, pp 1322–1328
Hospedales T, Antoniou A, Micaelli P, Storkey A (2021) Meta-learning in neural networks: a survey. IEEE Trans Pattern Anal Mach Intell 44(9):5149–5169
Huang S, Liu Q (2022) Addressing scale imbalance for small object detection with dense detector. Neurocomputing 473:68–78
Huang C, Li Y, Loy CC, Tang X (2019) Deep imbalanced learning for face recognition and attribute prediction. IEEE Trans Pattern Anal Mach Intell 42(11):2781–2794
Hung L-C, Hu Y-H, Tsai C-F, Huang M-W (2022) A dynamic time warping approach for handling class imbalanced medical datasets with missing values: a case study of protein localization site prediction. Expert Syst Appl 192:116437
Jamal MA, Brown M, Yang M-H, Wang L, Gong B (2020) Rethinking classbalanced methods for long-tailed visual recognition from a domain adaptation perspective. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 7610–7619
Jáñez-Martino F, Alaiz-Rodríguez R, Gonzále-Castro V, Fidalgo E, Alegre E (2023) A review of spam email detection: analysis of spammer strategies and the dataset shift problem. Artif Intell Rev 56(2):1145–1173
Jiang J, Liu F, Liu Y, Tang Q, Wang B, Zhong G, Wang W (2022) A dynamic ensemble algorithm for anomaly detection in iot imbalanced data streams. Comput Commun 194:250–257
Jiao B, Guo Y, Gong D, Chen Q (2022) Dynamic ensemble selection for imbalanced data streams with concept drift. IEEE Trans Neural Netw Learn Syst 1–14
Jin Q, Yuan M, Wang H, Wang M, Song Z (2022) Deep active learning models for imbalanced image classification. Knowl-Based Syst 257:109817
Kang Q, Chen X, Li S, Zhou M (2016) A noise-filtered under-sampling scheme for imbalanced classification. IEEE Trans Cybern 47(12):4263–4274
Kang Q, Shi L, Zhou M, Wang X, Wu Q, Wei Z (2017) A distance-based weighted undersampling scheme for support vector machines and its application to imbalanced classification. IEEE Trans Neural Netw Learn Syst 29(9):4152–4165
Kang B, Li Y, Xie S, Yuan Z, Feng J (2020) Exploring balanced feature spaces for representation learning. In: International conference on learning representations

Kang B, Xie S, Rohrbach M, Yan Z, Gordo A, Feng J, Kalantidis Y (2019) Decoupling representation and classifier for long-tailed recognition. arXiv:1910.09217
Karthik S, Revaud J, Chidlovskii B (2021) Learning from long-tailed data with noisy labels. arXiv:2108.11096
Kaur H, Pannu HS (2019) Malhi AK (2019) A systematic review on imbalanced data challenges in machine learning: applications and solutions. ACM Comput Surv (CSUR) 52(4):1–36
Kim KH, Sohn SY (2020) Hybrid neural network with cost-sensitive support vector machine for class-imbalanced multimodal data. Neural Netw 130:176–184
Kim M-J, Kang D-K, Kim HB (2015) Geometric mean based boosting algorithm with over-sampling to resolve data imbalance problem for bankruptcy prediction. Expert Syst Appl 42(3):1074–1082
Kim D, Yu H, Lee H, Beighley E, Durand M, Alsdorf DE, Hwang E (2019) Ensemble learning regression for estimating river discharges using satellite altimetry data: central congo river as a test-bed. Remote Sens Environ 221:741–755
Kim J, Hur Y, Park S, Yang E, Hwang SJ, Shin J (2020) Distribution aligning refinery of pseudo-label for imbalanced semisupervised learning. Adv Neural Inf Process Syst 33:14567–14579
Kim J, Jeong J, Shin J (2020) M2m: imbalanced classification via major-tominor translation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 13896–13905
Klikowski J, Woźniak M (2022) Deterministic sampling classifier with weighted bagging for drifted imbalanced data stream classification. Appl Soft Comput 122:108855
Korycki L, Krawczyk B (2021) Concept drift detection from multi-class imbalanced data streams. In: 2021 IEEE 37th International conference on data engineering (ICDE). IEEE, pp 1068–1079
Krawczyk B (2016) Learning from imbalanced data: open challenges and future directions. Progress in Artif Intell 5(4):221–232
Krawczyk B, Minku LL, Gama J, Stefanowski J, Woźniak M (2017) Ensemble learning for data stream analysis: a survey. Inf Fusion 37:132–156
Kuang J, Xu G, Tao T, Wu Q (2021) Classimbalance adversarial transfer learning network for cross-domain fault diagnosis with imbalanced data. IEEE Trans Instrum Meas 71:1–11
Lee HB, Lee H, Na D, Kim S, Park M, Yang E, Hwang SJ (2019) Learning to balance: Bayesian meta-learning for imbalanced and out-of-distribution tasks. arXiv:1905.12917
Lee H, Shin S, Kim H (2021) Abc: auxiliary balanced classifier for class-imbalanced semi-supervised learning. Adv Neural Inf Process Syst 34:7082–7094
Li L, He H, Li J (2019) Entropy-based sampling approaches for multi-class imbalanced problems. IEEE Trans Knowl Data Eng 32(11):2159–2170
Li Z, Huang W, Xiong Y, Ren S, Zhu T (2020) Incremental learning imbalanced data streams with concept drift: the dynamic updated ensemble algorithm. Knowl-Based Syst 195:105694
Li Z, Huang M, Liu G, Jiang C (2021) A hybrid method with dynamic weighted entropy for handling the problem of class imbalance with overlap in credit card fraud detection. Expert Syst Appl 175:114750
Li F, Liu S, Li K, Zhang Y, Duan M, Yao Z, Zhu G, Guo Y, Wang Y, Huang L et al (2023) Epiteamdna: sequence feature representation via transfer learning and ensemble learning for identifying multiple dna epigenetic modification types across species. Comput Biol Med 160:107030
Liang Z, Wang H, Yang K, Shi Y (2022) Adaptive fusion based method for imbalanced data classification. Front Neurorobot 16:827913
Lin W-C, Tsai C-F, Hu Y-H, Jhang J-S (2017) Clustering-based undersampling in class-imbalanced data. Inf Sci 409:17–26
Lin N, Fu S, Lin X, Wang L (2022b) Multi-label emotion classification based on adversarial multi-task learning. Inf Process Manag 59(6):103097
Lin N, Fu Y, Lin X, Yang A, Jiang S (2022) Cl-xabsa: contrastive learning for crosslingual aspect-based sentiment analysis. arXiv:2204.00791
Liu X-Y, Wu J, Zhou Z-H (2009) Exploratory undersampling for class-imbalance learning. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics) 39(2):539–550
Liu B, Wang S, Dong Q, Li S, Liu X (2016) Identification of dna-binding proteins by combining auto-cross covariance transformation and ensemble learning. IEEE Trans Nanobiosci 15(4):328–334
Liu Z, Wei P, Jiang J, Cao W, Bian J, Chang Y (2020) Mesa: boost ensemble imbalanced learning with meta-sampler. Adv Neural Inf Process Syst 33:14463–14474
Liu W, Zhang H, Ding Z, Liu Q, Zhu C (2021) A comprehensive active learning method for multiclass imbalanced data streams with concept drift. Knowl-Based Syst 215:106778
Liu H, Liu Z, Jia W, Zhang D, Tan J (2021) A novel imbalanced data classification method based on weakly supervised learning for fault diagnosis. IEEE Trans Industr Inf 18(3):1583–1593
Liu R, Liu Y, Duan J, Hou F, Wang L, Zhang X, Li G (2022) Ensemble learning directed classification and regression of hydrocarbon fuels. Fuel 324:124520

Liu Y, Yang G, Qiao S, Liu M, Qu L, Han N, Wu T, Yuan G, Peng Y (2023) Imbalanced data classification: using transfer learning and active sampling. Eng Appl Artif Intell 117:105621
Liu Z, Cao W, Gao Z, Bian J, Chen H, Chang Y, Liu T-Y (2020) Self-paced ensemble for highly imbalanced massive data classification. In: 2020 IEEE 36th international conference on data engineering (ICDE). IEEE, pp 841–852
Liu B, Li H, Kang H, Hua G, Vasconcelos N (2021) Gistnet: a geometric structure transfer network for long-tailed recognition. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 8209–8218
Li Z, Yu Z, Yang K, Shi Y, Xu Y, Chen CP (2021) Local tangent generative adversarial network for imbalanced data classification. In: 2021 International joint conference on neural networks (IJCNN). IEEE, pp 1–8
Longadge R, Dongre S (2013) Class imbalance problem in data mining review. arXiv:1305.1707
Lu J, Liu A, Dong F, Gu F, Gama J, Zhang G (2018) Learning under concept drift: a review. IEEE Trans Knowl Data Eng 31(12):2346–2363
Lu Y, Cheung Y-M, Tang YY (2019) Selfadaptive multiprototype-based competitive learning approach: a k-means-type algorithm for imbalanced data clustering. IEEE Trans Cybern 51(3):1598–1612
Lu Y, Cheung Y-M, Tang YY (2019) Adaptive chunk-based dynamic weighted majority for imbalanced data streams with concept drift. IEEE Trans Neural Netw Learn Syst 31(8):2764–2778
Lu Y, Cheung Y-m, Tang YY (2017) Dynamic weighted majority for incremental learning of imbalanced data streams with concept drift. In: IJCAI, pp 2393–2399
Lyu P, Zheng P, Yu W, Liu C, Xia M (2022) A novel multiview sampling-based meta self-paced learning approach for classimbalanced intelligent fault diagnosis. IEEE Trans Instrum Meas 71:1–12
Mani I, Zhang I (2003) knn approach to unbalanced data distributions: a case study involving information extraction. In: Proceedings of workshop on learning from imbalanced datasets, vol 126. ICML, pp 1–7
Mullick SS, Datta S, Das S (2019) Generative adversarial minority oversampling. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 1695–1704
Mullick SS, Datta S, Das S (2018) Adaptive learning-based k -nearest neighbor classifiers with resilience to class imbalance. IEEE Trans Neural Netw Learn Syst 29(11):5713–5725
Nam G, Jang S, Lee J (2023) Decoupled training for long-tailed classification with stochastic representations. arXiv:2304.09426
Ng WW, Zhang J, Lai CS, Pedrycz W, Lai LL, Wang X (2018) Cost-sensitive weighting and imbalance-reversed bagging for streaming imbalanced and concept drifting in electricity pricing classification. IEEE Trans Industr Inf 15(3):1588–1597
Nguwi Y-Y, Cho S-Y (2010) An unsupervised self-organizing learning with support vector ranking for imbalanced datasets. Expert Syst Appl 37(12):8303–8312
Oksuz K, Cam BC, Kalkan S, Akbas E (2020) Imbalance problems in object detection: a review. IEEE Trans Pattern Anal Mach Intell 43(10):3388–3415
Pan T, Zhao J, Wu W, Yang J (2020) Learning imbalanced datasets based on smote and gaussian distribution. Inf Sci 512:1214–1233
Pereira J, Saraiva F (2021) Convolutional neural network applied to detect electricity theft: a comparative study on unbalanced data handling techniques. Int J Electr Power Energy Syst 131:107085
Phung NM, Mimura M (2021) Detection of malicious javascript on an imbalanced dataset. Internet of Things 13:100357
Pourhabibi T, Ong K-L, Kam BH, Boo YL (2020) Fraud detection: a systematic literature review of graph-based anomaly detection approaches. Decis Support Syst 133:113303
Rana P, Sowmya A, Meijering E, Song Y (2023) Imbalanced classification for protein subcellular localization with multilabel oversampling. Bioinformatics 39(1):841
Rao S, Verma AK, Bhatia T (2023) Hybrid ensemble framework with self-attention mechanism for social spam detection on imbalanced data. Expert Syst Appl 217:119594
Razavi-Far R, Farajzadeh-Zanajni M, Wang B, Saif M, Chakrabarti S (2019) Imputation-based ensemble techniques for class imbalance learning. IEEE Trans Knowl Data Eng 33(5):1988–2001
Razavi-Far R, Farajzadeh-Zanajni M, Wang B, Saif M, Chakrabarti S (2021) Imputation-based ensemble techniques for class imbalance learning. IEEE Trans Knowl Data Eng 33(5):1988–2001
Ren S, Liao B, Zhu W, Li Z, Liu W, Li K (2018) The gradual resampling ensemble for mining imbalanced data streams with concept drift. Neurocomputing 286:150–166
Ren H, Wang J, Dai J, Zhu Z (2022) Liu J (2022) Dynamic balanced domain-adversarial networks for cross-domain fault diagnosis of train bearings. IEEE Trans Instrum Meas 71:1–12
Ren Z, Lin T, Feng K, Zhu Y, Liu Z, Yan K (2023) A systematic review on imbalanced learning methods in intelligent fault diagnosis. IEEE Trans Instrum Meas 72:1–35
Ren H, Tang Y, Dong W, Ren S, Jiang L (2023) Duen: dynamic ensemble handling class imbalance in network intrusion detection. Expert Syst Appl 229:120420

Ren J, Zhang M, Yu C, Liu Z (2022) Balanced mse for imbalanced visual regression. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 7926–7935
Rezvani S, Wang X (2023) A broad review on class imbalance learning techniques. Appl Soft Comput 110415
Sáez JA, Luengo J, Stefanowski J, Herrera F (2015) Smote-ipf: addressing the noisy and borderline examples problem in imbalanced classification by a re-sampling method with filtering. Inf Sci 291:184–203
Sağlam F, Cengiz MA (2022) A novel smotebased resampling technique trough noise detection and the boosting procedure. Expert Syst Appl 200:117023
Sahani M, Dash PK (2019) Fpga-based online power quality disturbances monitoring using reduced-sample hht and class-specific weighted rvfln. IEEE Trans Industr Inf 15(8):4614–4623
Saini M, Susan S (2020) Deep transfer with minority data augmentation for imbalanced breast cancer dataset. Appl Soft Comput 97:106759
Saini M, Susan S (2022) Vggin-net: deep transfer network for imbalanced breast cancer dataset. IEEE/ACM Trans Comput Biol Bioinf 20(1):752–762
Samuel D, Chechik G (2021) Distributional robustness loss for long-tail learning. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 9495–9504
Seiffert C, Khoshgoftaar TM, Van Hulse J, Napolitano A (2009) Rusboost: a hybrid approach to alleviating class imbalance. IEEE transactions on systems, man, and cybernetics-part A: systems and humans 40(1):185–197
Sharma T (2021) Rattan D (2021) Malicious application detection in android—a systematic literature review. Comput Sci Rev 40:100373
Shorten C, Khoshgoftaar TM (2019) A survey on image data augmentation for deep learning. J Big Data 6(1):1–48
Singh R, Ahmed T, Kumar A, Singh AK, Pandey AK, Singh SK (2020) Imbalanced breast cancer classification using transfer learning. IEEE/ACM Trans Comput Biol Bioinf 18(1):83–93
Sleeman WC IV, Kapoor R, Ghosh P (2022) Multimodal classification: current landscape, taxonomy and future directions. ACM Comput Surv 55(7):1–31
Smith MR, Martinez T, Giraud-Carrier C (2014) An instance level analysis of data complexity. Mach Learn 95:225–256
Stefanowski J, Wilk S (2008) Selective preprocessing of imbalanced data for improving classification performance. In: Data warehousing and knowledge discovery: 10th international conference, DaWaK 2008 Turin, Italy, September 2-5, 2008 Proceedings 10. Springer, pp 283–292
Steininger M, Kobs K, Davidson P, Krause A, Hotho A (2021) Density-based weighting for imbalanced regression. Mach Learn 110:2187–2211
Sun J, Lang J, Fujita H, Li H (2018) Imbalanced enterprise credit evaluation with dte-sbd: decision tree ensemble based on smote and bagging with differentiated sampling rates. Inf Sci 425:76–91
Sun Y, Cai L, Liao B, Zhu W (2020) Minority sub-region estimation-based oversampling for imbalance learning. IEEE Trans Knowl Data Eng 34(5):2324–2334
Sun Y, Cai L, Liao B, Zhu W, Xu J (2022) A robust oversampling approach for class imbalance problem with small disjuncts. IEEE Trans Knowl Data Eng
Tang K, Huang J, Zhang H (2020) Longtailed classification by keeping the good and removing the bad momentum causal effect. Adv Neural Inf Process Syst 33:1513–1524
Tarekegn AN, Giacobini M, Michalak K (2021) A review of methods for imbalanced multi-label classification. Pattern Recogn 118:107965
Torgo L, Ribeiro R (2009) Precision and recall for regression. In: Discovery science: 12th international conference, DS 2009, Porto, Portugal, October 3-5, 2009 12. Springer, pp 332–346
Tsai C-F, Lin W-C, Hu Y-H, Yao G-T (2019) Under-sampling class imbalanced datasets by combining clustering analysis and instance selection. Inf Sci 477:47–54
Van Hulse J, Khoshgoftaar T (2009) Knowledge discovery from imbalanced and noisy data. Data Knowl Eng 68(12):1513–1542
Viola P, Jones M (2001) Fast and robust classification using asymmetric adaboost and a detector cascade. Adv Neural Inf Process Syst 14
Wang B, Pineau J (2016) Online bagging and boosting for imbalanced data streams. IEEE Trans Knowl Data Eng 28(12):3353–3366
Wang F, Wei L (2022) Multi-scale deep learning for the imbalanced multi-label protein subcellular localization prediction based on immunohistochemistry images. Bioinformatics 38(9):2602–2611
Wang C, Hu L, Guo M, Liu X, Zou Q (2015) imdc: an ensemble learning method for imbalanced classification with mirna data. Genet Mol Res 14(1):123–133
Wang Y, Yao Q, Kwok JT, Ni LM (2020) Generalizing from a few examples: a urvey on few-shot learning. ACM Comput Surv (csur) 53(3):1–34

Wang Z, Cao C, Zhu Y (2020) Entropy and confidence-based undersampling boosting random forests for imbalanced problems. IEEE Trans Neural Netw Learn Syst 31(12):5178–5191
Wang L, Zhang L, Qi X, Yi Z (2021) Deep attention-based imbalanced image classification. IEEE Trans Neural Netw Learn Syst 33(8):3320–3330
Wang Y, Gan W, Yang J, Wu W, Yan J (2019) Dynamic curriculum learning for imbalanced data classification. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 5017–5026
Wang P, Han K, Wei X-S, Zhang L, Wang L (2021) Contrastive learning based hybrid networks for long-tailed image classification. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 943–952
Wang X, Lian L, Miao Z, Liu Z, Yu SX (2020) Long-tailed recognition by routing diverse distribution-aware experts. arXiv:2010.01809
Wang L, Xu S, Wang X, Zhu Q (2021) Addressing class imbalance in federated learning. In: Proceedings of the AAAI conference on artificial intelligence, vol 35, pp 10165–10173
Wang S, Yao X (2009) Diversity analysis on imbalanced data sets by using ensemble models. In: 2009 IEEE symposium on computational intelligence and data mining. IEEE, pp 324–331
Wei C, Sohn K, Mellina C, Yuille A, Yang F (2021) Crest: a class-rebalancing selftraining framework for imbalanced semisupervised learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 10857–10866
Wen G, Li X, Zhu Y, Chen L, Luo Q, Tan M (2021) One-step spectral rotation clustering for imbalanced high-dimensional data. Inf Process Manag 58(1):102388
Wilson DL (1972) Asymptotic properties of nearest neighbor rules using edited data. IEEE Trans Syst Man Cybern (3):408–421
Woźniak M, Grana M, Corchado E (2014) A survey of multiple classifier systems as hybrid systems. Information Fusion 16:3–17
Wu T-Y, Morgado P, Wang P, Ho C-H, Vasconcelos N (2020) Solving long-tailed recognition with deep realistic taxonomic classifier. In: Computer vision-ECCV 2020: 16th European conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part VIII 16. Springer, pp 171–189
Wu X, Meng S (2016) E-commerce customer churn prediction based on improved smote and adaboost. In: 2016 13th International conference on service systems and service management (ICSSSM). IEEE, pp 1–5
Xiang L, Ding G, Han J (2020) Learning from multiple experts: self-paced knowledge distillation for long-tailed classification. In: Computer vision-ECCV 2020: 16th European conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part V 16. Springer, pp 247–263
Xiao Y, Wu J, Lin Z (2021) Cancer diagnosis using generative adversarial networks based on deep learning from imbalanced data. Comput Biol Med 135:104540
Xu Y, Yu Z, Chen CP, Liu Z (2021) Adaptive subspace optimization ensemble method for high-dimensional imbalanced data classification. IEEE Trans Neural Netw Learn Syst
Yan Z, Wen H (2021) Electricity theft detection base on extreme gradient boosting in ami. IEEE Trans Instrum Meas 70:1–9
Yan Y, Zhu Y, Liu R, Zhang Y, Zhang Y, Zhang L (2023) Spatial distribution-based imbalanced undersampling. IEEE Trans Knowl Data Eng 35(6):6376–6391
Yang Y, Xu Z (2020) Rethinking the value of labels for improving class-imbalanced learning. Adv Neural Inf Process Syst 33:19290–19301
Yang K, Yu Z, Wen X, Cao W, Chen CP, Wong H-S, You J (2019) Hybrid classifier ensemble for imbalanced data. IEEE Trans Neural Netw Learn Syst 31(4):1387–1400
Yang K, Yu Z, Chen CP, Cao W, Wong H-S, You J, Han G (2021) Progressive hybrid classifier ensemble for imbalanced data. IEEE Trans Syst, Man, and Cybernetics: Systems 52(4):2464–2478
Yang K, Yu Z, Chen CP, Cao W, You J, Wong H-S (2021) Incremental weighted ensemble broad learning system for imbalanced data. IEEE Trans Knowl Data Eng 34(12):5809–5824
Yang K, Shi Y, Yu Z, Yang Q, Sangaiah AK, Zeng H (2022) Stacked one-class broad learning system for intrusion detection in industry 4.0. IEEE Trans Ind Inform 19(1):251–260
Yang Z, Liu X, Li T, Wu D, Wang J, Zhao Y, Han H (2022) A systematic literature review of methods and datasets for anomaly-based network intrusion detection. Comput Secur 116:102675
Yang K, Chen W, Bi J, Wang M, Luo F (2023) Multi-view broad learning system for electricity theft detection. Appl Energy 352:121914
Yang Y, Lv H, Chen N (2023) A survey on ensemble learning under the era of deep learning. Artif Intell Rev 56(6):5545–5589
Yang Y, Zha K, Chen Y, Wang H, Katabi D (2021) Delving into deep imbalanced regression. In: International conference on machine learning. PMLR, pp 11842–11851

Yan Z, Hongle D, Gang K, Lin Z, Chen Y-C (2022) Dynamic weighted selective ensemble learning algorithm for imbalanced data streams. J Supercomput 1–26
Yin L, Du X, Ma C, Gu H (2022) Virtual screening of drug proteins based on the prediction classification model of imbalanced data mining. Processes 10(7):1420
You D, Xiao J, Wang Y, Yan H, Wu D, Chen Z, Shen L, Wu X (2023) Online learning from incomplete and imbalanced data streams. IEEE Trans Knowl Data Eng
Zang Y, Huang C, Loy CC (2021) Fasa: feature augmentation and sampling adaptation for long-tailed instance segmentation. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 3457–3466
Zhang X, Hu B-G (2014) A new strategy of cost-free learning in the class imbalance problem. IEEE Trans Knowl Data Eng 26(12):2872–2885
Zhang H, Li M (2014) Rwo-sampling: a random walk over-sampling approach to imbalanced data classification. Inf Fusion 20:99–116
Zhang T, Ma F, Yue D, Peng C, O'Hare GM (2019) Interval type-2 fuzzy local enhancement based rough k-means clustering considering imbalanced clusters. IEEE Trans Fuzzy Syst 28(9):1925–1939
Zhang H, Liu W, Liu Q (2020) Reinforcement online active learning ensemble for drifting imbalanced data streams. IEEE Trans Knowl Data Eng 34(8):3971–3983
Zhang T, Chen J, Li F, Zhang K, Lv H, He S, Xu E (2022) Intelligent fault diagnosis of machines with small & imbalanced data: a state-of-the-art review and possible extensions. ISA Trans 119:152–171
Zhang Z, Wang G, Carranza EJM, Fan J, Liu X, Zhang X, Dong Y, Chang X, Sha D (2022) An integrated framework for datadriven mineral prospectivity mapping using bagging-based positive-unlabeled learning and bayesian cost-sensitive logistic regression. Nat Resour Res 31(6):3041–3060
Zhang Y, Kang B, Hooi B, Yan S, Feng J (2023) Deep long-tailed learning: a survey. IEEE Trans Pattern Anal Mach Intell
Zhang J, Tao H, Hou C (2023) Imbalanced clustering with theoretical learning bounds. IEEE Trans Knowl Data Eng
Zhang X, Wu Z, Weng Z, Fu H, Chen J, Jiang Y-G, Davis LS (2021) Videolt: largescale long-tailed video recognition. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 7960–7969
Zhang Y, Zhang H, Lin Y (2022) Data augmentation for long-tailed and imbalanced polyphone disambiguation in mandarin. In: ICASSP 2022-2022 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, pp 7137–7141
Zhauniarovich Y, Khalil I, Yu T, Dacier M (2018) A survey on malicious domains detection through dns data analysis. ACM Comput Surv (CSUR) 51(4):1–36
Zhou H, Zhang J, Luo T, Yang Y, Lei J (2022) Debiased scene graph generation for dual imbalance learning. IEEE Trans Pattern Anal Mach Intell 45(4):4274–4288
Zhou B, Cui Q, Wei X-S, Chen Z-M (2020) Bbn: bilateral-branch network with cumulative learning for long-tailed visual recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 9719–9728
Zhu T, Lin Y, Liu Y, Zhang W, Zhang J (2019) Minority oversampling for imbalanced ordinal regression. Knowl-Based Syst 166:140–155
Zhu H, Zhou M, Liu G, Xie Y, Liu S, Guo C (2023) Nus: noisy-sample-removed undersampling scheme for imbalanced classification and application to credit card fraud detection. IEEE Trans Comput Soc Syst
Zyblewski P, Sabourin R, Woźniak M (2021) Preprocessed dynamic classifier ensemble selection for highly imbalanced drifted data streams. Inf Fusion 66:138–154