INTRODUCTION

HAIBO HE
Department of Electrical, Computer, and Biomedical Engineering, University of Rhode Island, Kingston, RI, USA

Abstract: With the continuous expansion of data availability in many large-scale, complex, and networked systems, it becomes critical to advance raw data from fundamental research on the Big Data challenge to support decision-making processes. Although existing machine-learning and data-mining techniques have shown great success in many real-world applications, learning from imbalanced data is a relatively new challenge. This book is dedicated to the state-of-the-art research on imbalanced learning, with a broader discussions on the imbalanced learning foundations, algorithms, databases, assessment metrics, and applications. In this chapter, we provide an introduction to problem formulation, a brief summary of the major categories of imbalanced learning methods, and an overview of the challenges and opportunities in this field. This chapter lays the structural foundation of this book and directs readers to the interesting topics discussed in subsequent chapters.

1.1 PROBLEM FORMULATION

We start with the definition of imbalanced learning in this chapter to lay the foundation for further discussions in the book. Specifically, we define imbalanced learning as the learning process for data representation and information extraction with severe data distribution skews to develop effective decision boundaries to support the decision-making process. The learning process could involve supervised learning, unsupervised learning, semi-supervised learning, or a combination
of two or all of them. The task of imbalanced learning could also be applied to regression, classification, or clustering tasks. In this Chapter, we provide a brief introduction to the problem formulation, research methods, and challenges and opportunities in this field. This chapter is based on a recent comprehensive survey and critical review of imbalanced learning as presented in [1], and interested readers could refer to that survey paper for more details regarding imbalanced learning.

Imbalanced learning not only presents significant new challenges to the data research community but also raises many critical questions in real-world data-intensive applications, ranging from civilian applications such as financial and biomedical data analysis to security- and defense-related applications such as surveillance and military data analysis [1]. This increased interest in imbalanced learning is reflected in the recent significantly increased number of publications in this field as well as in the organization of dedicated workshops, conferences, symposiums, and special issues, [2, 3, 4].

To start with a simple example of imbalanced learning, let us consider a popular case study in biomedical data analysis [1]. Consider the “Mammography Data Set,” a collection of images acquired from a series of mammography examinations performed on a set of distinct patients [5–7]. For such a dataset, the natural classes that arise are “Positive” or “Negative” for an image representative of a “cancerous” or “healthy” patient, respectively. From experience, one would expect the number of noncancerous patients to exceed greatly the number of cancerous patients; indeed, this dataset contains 10,923 “Negative” (majority class) and 260 “Positive” (minority class) samples. Preferably, we require a classifier that provides a balanced degree of predictive accuracy for both the minority and majority classes on the dataset. However, in many standard learning algorithms, we find that classifiers tend to provide a severely imbalanced degree of accuracy, with the majority class having close to 100% accuracy and the minority class having accuracies of 0 – 10%; see for instance, [5, 7]. Suppose a classifier achieves 5% accuracy on the minority class of the mammography dataset. Analytically, this would suggest that 247 minority samples are misclassified as majority samples (i.e., 247 cancerous patients are diagnosed as noncancerous). In the medical industry, the ramifications of such a consequence can be overwhelmingly costly, more so than classifying a noncancerous patient as cancerous [8]. Furthermore, this also suggests that the conventional evaluation practice of using singular assessment criteria, such as the overall accuracy or error rate, does not provide adequate information in the case of imbalanced learning. In an extreme case, if a given dataset includes 1% of minority class examples and 99% of majority class examples, a naive approach of classifying every example to be a majority class example would provide an accuracy of 99%. Taken at face value, 99% accuracy across the entire dataset appears superb; however, by the same token, this description fails to reflect the fact that none of the minority examples are identified, when in many situations, those minority examples are of much more interest. This clearly demonstrates the need to revisit the assessment metrics for imbalanced learning, which is discussed in Chapter 8.
STATE-OF-THE-ART RESEARCH

1.2 STATE-OF-THE-ART RESEARCH

Given the new challenges facing imbalanced learning, extensive efforts and significant progress have been made in the community to tackle this problem. In this section, we provide a brief summary of the major category of approaches for imbalanced learning. Our goal is just to highlight some of the major research methodologies while directing the readers to different chapters in this book for the latest research development in each category of approach. Furthermore, a comprehensive summary and critical review of various types of imbalanced learning techniques can also be found in a recent survey [1].

1.2.1 Sampling Methods

Sampling methods seem to be the dominate type of approach in the community as they tackle imbalanced learning in a straightforward manner. In general, the use of sampling methods in imbalanced learning consists of the modification of an imbalanced dataset by some mechanism in order to provide a balanced distribution. Representative work in this area includes random oversampling [9], random undersampling [10], synthetic sampling with data generation [5, 11–13], cluster-based sampling methods [14], and integration of sampling and boosting [6, 15, 16].

The key aspect of sampling methods is the mechanism used to sample the original dataset. Under different assumptions and with different objective considerations, various approaches have been proposed. For instance, the mechanism of random oversampling follows naturally from its description by replicating a randomly selected set of examples from the minority class. On the basis of such simple sampling techniques, many informed sampling methods have been proposed, such as the EasyEnsemble and BalanceCascade algorithms [17]. Synthetic sampling with data generation techniques has also attracted much attention. For example, the synthetic minority oversampling technique (SMOTE) algorithm creates artificial data based on the feature space similarities between existing minority examples [5]. Adaptive sampling methods have also been proposed, such as the borderline-SMOTE [11] and adaptive synthetic (ADASYN) sampling [12] algorithms. Sampling strategies have also been integrated with ensemble learning techniques by the community, such as in SMOTEBoost [15], RAMOBoost [18], and DataBoost-IM [6]. Data-cleaning techniques, such as Tomek links [19], have been effectively applied to remove the overlapping that is introduced from sampling methods for imbalanced learning. Some representative work in this area includes the one-side selection (OSS) method [13] and the neighborhood cleaning rule (NCL) [20].

1.2.2 Cost-Sensitive Methods

Cost-sensitive learning methods target the problem of imbalanced learning by using different cost matrices that describe the costs for misclassifying any
particular data example [21, 22]. Research in the past indicates that there is a strong connection between cost-sensitive learning and imbalanced learning [4, 23, 24]. In general, there are three categories of approaches to implement cost-sensitive learning for imbalanced data. The first class of techniques applies misclassification costs to the dataset as a form of dataspace weighting (translation theorem [25]); these techniques are essentially cost-sensitive bootstrap sampling approaches where misclassification costs are used to select the best training distribution. The second class applies cost-minimizing techniques to the combination schemes of ensemble methods (Metacost framework [26]); this class consists of various meta techniques, such as the AdaC1, AdaC2, and AdaC3 methods [27] and AdaCost [28]. The third class of techniques incorporates cost-sensitive functions or features directly into classification paradigms to essentially “fit” the cost-sensitive framework into these classifiers, such as the cost-sensitive decision trees [21, 24], cost-sensitive neural networks [29, 30], cost-sensitive Bayesian classifiers [31, 32], and cost-sensitive support vector machines (SVMs) [33–35].

1.2.3 Kernel-Based Learning Methods

There have been many studies that integrate kernel-based learning methods with general sampling and ensemble techniques for imbalanced learning. Some examples include the SMOTE with different costs (SDC) method [36] and the ensembles of over/undersampled SVMs [37, 38]. For example, the SDC algorithm uses different error costs [36] for different classes to bias the SVM to guarantee a more well-defined boundary. The granular support vector machines—repetitive undersampling (GSVM-RU) algorithm was proposed in [39] to integrate SVM learning with undersampling methods. Another major category of kernel-based learning research efforts focuses more concretely on the mechanisms of the SVM itself; this group of methods are often called kernel modification methods, such as the kernel classifier construction algorithm proposed in [40]. Other examples of kernel modification include the various techniques used for adjusting the SVM class boundary [41, 42]. Furthermore, the total margin-based adaptive fuzzy SVM (TAF-SVM) kernel method was proposed in [43] to improve SVM robustness. Other major kernel modification methods include the $k$-category proximal SVM (PSVM) [44], SVMs for extreme imbalanced datasets [45], support cluster machines (SCMs) [46], kernel neural gas (KNG) algorithm [47], hybrid kernel machine ensemble (HKME) algorithm [48], and the Adaboost relevance vector machine (RVM) [49].

1.2.4 Active Learning Methods

Active learning methods have also been proposed for imbalanced learning in the literature [50–53]. For instance, Ertekin et al. [51, 52] proposed an efficient SVM-based active learning method that queries a small pool of data at each
iterative step of active learning instead of querying the entire dataset. Active learning integrations with sampling techniques have also been proposed. For instance, Zhu and Hovy [54] analyzed the effect of undersampling and oversampling techniques with active learning for the word sense disambiguation (WSD) imbalanced learning problem. Another active learning sampling method is the simple active learning heuristic (SALH) approach proposed in [55]. The main aim of this method is to provide a generic model for the evolution of genetic programming (GP) classifiers by integrating the stochastic subsampling method and a modified Wilcoxon–Mann–Whitney (WMW) cost function [55]. Major advantages of the SALH method include the ability to actively bias the data distribution for learning, the existence of a robust cost function, and the improvement of the computational cost related to the fitness evaluation.

1.2.5 One-Class Learning Methods

The one-class learning or novelty detection method has also attracted much attention in the community for imbalanced learning [4]. Generally speaking, this category of approaches aims to recognize instances of a concept by using mainly, or only, a single class of examples (i.e., recognition-based methodology) rather than differentiating between instances of both positive and negative classes as in the conventional learning approaches (i.e., discrimination-based inductive methodology). Representative work in this area includes the one-class SVMs [56, 57] and the autoassociator (or autoencoder) method [58–60]. For instance, in [59], a comparison between different sampling methods and the one-class autoassociator method was presented. The novelty detection approach based on redundancy compression and nonredundancy differentiation techniques was investigated in [60]. Lee and Cho [61] suggested that novelty detection methods are particularly useful for extremely imbalanced datasets, whereas regular discrimination-based inductive classifiers are suitable for relatively moderate imbalanced datasets.

Although the current efforts in the community are focused on two-class imbalanced problems, multi-class imbalanced learning problems also exist and have been investigated in numerous works. For instance, in [62], a cost-sensitive boosting algorithm AdaC2.M1 was proposed to tackle the class imbalance problem with multiple classes. In [63], an iterative method for multi-class cost-sensitive learning was proposed. Other works of multi-class imbalanced learning include the min–max modular network [64] and the rescaling approach for multi-class cost-sensitive neural networks [65], to name a few.

Our discussions in this section by no means provide a full coverage of the complete set of methods to tackle the imbalanced learning problem, given the variety of assumptions for the imbalanced data and different learning objectives of different applications. Interested readers can refer to [1] for a recent survey of the imbalanced learning methods. The latest research development on this topic can be found in the following chapters.
1.3 LOOKING AHEAD: CHALLENGES AND OPPORTUNITIES

Given the increased complexity of data in many of the current real-world applications, imbalanced learning presents many new challenges as well as opportunities to the community. Here, we highlight a few of those to hopefully provide some suggestions for long-term research in this field.

1.3.1 Advancement of the Foundations and Principles of Imbalanced Learning

Although there are numerous new efforts in the community targeting imbalanced learning, many of the current research methodologies are very heuristic and ad hoc based. There is a lack of theoretical foundation and principles to guide the development of systematic imbalanced learning approaches. For instance, although almost every paper presented to the community claims that there is a certain degree of improvement on learning performance or efficiency, there are situations where learning from the original imbalanced data could provide better performance. This raises important questions: what is the assurance that algorithms specifically designed for imbalanced learning could really help, and how and why? [1]. Can one simply design robust-enough algorithms that could learn from whatever data are presented [53]? Also, as there are many existing base learning algorithms such as the decision tree, neural network, and SVM, is there a way we could develop a theoretical guidance on which base learning algorithm is most appropriate for a particular type of imbalanced data? Are there any error bounds for those base learning algorithms for imbalanced data? What is the relationship between data-imbalanced ratio and learning model complexity? What are the best levels of balanced ratio for a given base learning algorithm? All of these are open questions to the community now. In fact, a thorough understanding of these questions will not only provide fundamental insights into the imbalanced learning problem but also provide critical technical tools and solutions to many practical real imbalanced learning applications. Therefore, it is essential for the community to investigate all, or at least some, of these questions for the long-term sustainable development of this field. In Chapter 2, a dedicated section discusses the foundations of imbalanced learning to provide some new insights along this direction.

1.3.2 Unified Data Benchmark for Imbalanced Learning

It is well known that data plays a key role in any kind of machine-learning and data-mining research. This is especially the case for the relatively new field of imbalanced learning. Although there are currently many publicly available benchmarks for assessing the effectiveness of different learning algorithm/tools (e.g., University of California, Irvine (UCI) data repository [66], and National
Institute of Standards and Technology (NIST) Scientific and Technical Databases [67]), there are very few data benchmarks that are solely dedicated to imbalanced learning problems. This has caused data for imbalanced learning to be very costly in the society. For instance, many of the existing data benchmarks require additional manipulation before they can be applied to imbalanced learning scenarios for each algorithm. This limitation has created a bottleneck in the long-term development of research in this field. Therefore, unified data benchmarks for imbalanced learning are important to provide an open-access source for the community not only to promote data sharing but also to provide a common platform to ensure a fair comparative study among different methods.

1.3.3 Standardized Assessment Metrics

As discussed in Section 1.1, traditional assessment techniques may not be able to provide a fair and comprehensive evaluation of the imbalanced learning algorithms. In particular, it is widely agreed that a singular evaluation metric, such as overall classification error rate, is not sufficient when handling imbalanced learning problems. As suggested in [1], it seems that a combination of singular-based metrics (e.g., precision, recall, $F$-measure, and G-mean) together with curve-based assessment metrics (e.g., receiver operating characteristic (ROC) curve, precision-recall (PR) curve, and cost curve) will provide a more complete assessment of imbalanced learning. Therefore, it is necessary for the community to establish—as a standard—the practice of using such assessment approaches to provide more insights into the advantages and limitations of different types of imbalanced learning methods. More details on this can be found in Chapter 8.

1.3.4 Emerging Applications with Imbalanced Learning

Imbalanced learning has presented itself to be an essential part in many critical real-world applications. For instance, in the aforementioned biomedical diagnosis situation, an effective learning approach that could handle the imbalanced data is key to supporting the medical decision-making process. Similar scenarios have appeared in many other mission-critical tasks, such as security (e.g., abnormal behavior recognition), defense (e.g., military data analysis), and financial industry (e.g., outlier detection). This book also presents a few examples of such critical applications to demonstrate the importance of imbalanced learning.

1.4 ACKNOWLEDGMENTS

This work was supported in part by the National Science Foundation (NSF) under grant ECCS 1053717 and Army Research Office (ARO) under Grant W911NF-12-1-0378.
REFERENCES

REFERENCES


