Hindawi
Journal of Healthcare Engineering
Volume 2018, Article ID 6275435, 11 pages
https://doi.org/10.1155/2018/6275435

Research Article

A Framework of Rebalancing Imbalanced Healthcare Data for Rare Events' Classification: A Case of Look-Alike Sound-Alike Mix-Up Incident Detection

Yang Zhao,[1] Zoie Shui-Yee Wong,[2] and Kwok Leung Tsui[1]

[1]*Department of Systems Engineering and Engineering Management, City University of Hong Kong, Kowloon, Hong Kong*
[2]*Graduate School of Public Health, St. Luke's International University, Tokyo, Japan*

Correspondence should be addressed to Zoie Shui-Yee Wong; zoiewong@luke.ac.jp

Received 11 September 2017; Revised 2 February 2018; Accepted 22 February 2018; Published 22 May 2018

Academic Editor: John S. Katsanis

Identifying rare but significant healthcare events in massive unstructured datasets has become a common task in healthcare data analytics. However, imbalanced class distribution in many practical datasets greatly hampers the detection of rare events, as most classification methods implicitly assume an equal occurrence of classes and are designed to maximize the overall classification accuracy. In this study, we develop a framework for learning healthcare data with imbalanced distribution via incorporating different rebalancing strategies. The evaluation results showed that the developed framework can significantly improve the detection accuracy of medical incidents due to look-alike sound-alike (LASA) mix-ups. Specifically, logistic regression combined with the synthetic minority oversampling technique (SMOTE) produces the best detection results, with a significant 45.3% increase in recall (recall = 75.7%) compared with pure logistic regression (recall = 52.1%).

1. Introduction

The rapid growth of electronic health records (EHRs) is generating massive health informatics and bioinformatics datasets, and more and more crowdsourced medical data are becoming available. Using statistical data analytics to detect rare but significant healthcare events in these massive unstructured dataset, such as medication errors and disease risk, has the potential to reduce treatment costs, avoid preventable diseases, and improve care quality in general [1, 2]. One major challenge to effective healthcare data analytics is highly skewed data class distribution, which is referred to as the imbalanced classification problem. An imbalanced classification problem occurs when the classes in a dataset have a highly unequal number of samples. For example, in a binary classification, the imbalanced classification problem is present when one class has significantly fewer observations than the other class. The former is usually called a minority class, and the latter, a majority class. In this study, we develop a method for detecting relevant healthcare events in datasets where this data challenge is present.

In some healthcare-related datasets with imbalanced classification, accurately detecting minority class observations is of great importance, as they correspond to high-impact events. For instance, some attempts have been made to automatically identify medical incident reports. The targeted medical incident reports are usually reports of incidents that have been recognized as common causes of medication errors that may result in adverse or harmful patient outcomes [3]. In practice, many datasets of medical incident reports exhibit imbalanced class distribution. For example, in an investigation of the classification of two types of medical incident reports, namely, "clinical management/inadequate handover" and "clinical management/incorrect patient," Ong (2010) [4] found that there were more than twice as many clinical management/incorrect patient cases as clinical management/inadequate handover cases. In another example, Wong (2016) [5] examined the detection

of look-alike and sound-alike (LASA) mix-up cases from the medical incident reports and found that only 21% of the available reports were related to LASA cases.

Conventional statistical learning classifiers typically perform poorly in imbalanced datasets, as they implicitly assume an equal occurrence of all classes and are designed to maximize the overall classification accuracy. Thus, these classifiers favor the majority class, resulting in poor accuracy in detecting minority class observations [6, 7]. Many healthcare data analytics applications have neglected the problem of dataset imbalance [8], and the effectiveness of classifiers that use rebalancing strategies to address the detection problem has rarely been evaluated [9, 10].

Resampling and cost-sensitive learning are state-of-the-art rebalancing strategies for imbalanced classification. The resampling schemes include randomly oversampling the minority class, undersampling the majority class, and some advanced synthetic sampling methods that attempt to rebalance class distribution at the data level. However, these rebalancing strategies have some limitations. For instance, an unavoidable consequence of undersampling is the loss of information [11], whereas oversampling through the random replication of the minority class sample usually creates very specific rules, leading to overfitting [7]. Cost-sensitive learning considers the costs of misclassified instances and minimizes the total misclassification cost, attempting to rebalance class distribution at the algorithm level. As cost-sensitive learning methods are motivated by the observation that most real applications do not have a unified cost for misclassification, the cost matrix needs be manually determined beforehand.

In this study, we develop a framework for analyzing healthcare data with imbalanced distribution that incorporates different rebalancing strategies, and we offer guidelines for choosing appropriate procedures. This learning framework consists of two main stages: selecting base classifiers and evaluating rebalancing strategies. We examine the effect of data imbalance on classifier performance and the effectiveness of various rebalancing strategies. The results of our analysis of a published study's dataset show that the developed framework significantly improves the accuracy in detecting medical incidents caused by LASA mix-ups. It is worth noting that the framework has a broad range of applications beyond medical incident reports detection, in datasets with similar imbalanced data properties.

2. Background

2.1. Imbalanced Data in Healthcare. The imbalance property that is common to many real healthcare datasets makes classification a challenging task. The imbalanced classification problem in the healthcare domain, where data are often highly skewed due to individual heterogeneity and diversity, affects issues such as cancer diagnostics [12, 13], patient safety informatics [5, 14], and disease risk prediction [15]. Most standard classifiers, such as logistic regression and the support vector machine, implicitly assume that both classes are equally common. Additionally, these methods are designed for maximizing overall classification accuracy. As a result, they favor the majority class, resulting in poor sensitivity toward the minority class [6, 7]. This intuition is illustrated in Figure 1, which contains a synthetic example containing a majority class and a minority class. The solid line (ω^*) depicts the optimal separator in the underlying distribution, and the dotted line ($\hat{\omega}$) is the max-margin loss-minimizing separator generated over the instances. In this case, the induced separator is clearly skewed toward the minority class.

The fundamental concern raised by the imbalanced learning problem is that the performance of most standard learning algorithms is significantly compromised by imbalanced data. Resampling at the data level and cost-sensitive learning at the algorithm level are common strategies for addressing imbalanced classification. In the following sections, we review these strategies and examine their effectiveness when they are combined with standard classifiers to detect medical incidents in imbalanced datasets.

2.2. Rebalancing Strategies

2.2.1. Data-Level Approaches. Approaches at the data level, based on the observation that classifiers learn better from a balanced distribution than from an imbalanced one, use various methods for rebalancing the class distribution [4, 16]. The representative scheme is randomly oversampling the minority class and undersampling the majority class [17, 18] or a combination of both schemes. Some advanced methods, called synthetic sampling strategies, generate synthetic instances to improve classifiers' performance.

(1) Oversampling and Undersampling. Due to their simplicity and computational efficiency, over- and undersampling methodologies are popular strategies for countering the effect of imbalanced datasets [19–24]. The oversampling technique consists of randomly selecting instances from the minority class with replication and then adding the replications into the minority class. In this way, the size of the minority class is enlarged. Let $\boldsymbol{X} = \{x_i\}$ $(i = 1, 2, \ldots, N)$ denote the minority class with N instances. Then, the randomly oversampled minority class is $\boldsymbol{X_o} = \{x_j\}$ $(j = 1, 2, \ldots, N_o)$, where $\forall x_j \in \boldsymbol{X}$ and $N_o > N$. In contrast, the undersampling scheme randomly removes instances from the majority class, which can also rebalance the minority and majority classes. Let $\boldsymbol{Z} = \{\boldsymbol{z_i}\}$ $(i = 1, 2, \ldots, M)$ denote the majority class with M instances. Then, the randomly undersampled majority class is $\boldsymbol{Z_o} = \{z_j\}$ $(j = 1, 2, \ldots, M_u)$, where $\forall z_j \in \boldsymbol{Z}$ and $M_u < M$. The random under- and oversampling methods each have shortcomings. An unavoidable drawback of undersampling is the loss of information [11], whereas oversampling, through the random replication of the minority class sample, usually creates very specific rules, leading to model overfitting [7].

(2) Synthetic Sampling. The synthetic minority oversampling technique (SMOTE) [25] is a typical synthetic sampling method that has been very successful in various applications and been the foundation for many variants of the basic synthetic sampling scheme [26]. SMOTE searches k-nearest minority neighbors of each minority instance (denoted as x_i) and then randomly selects one of the neighbors as the

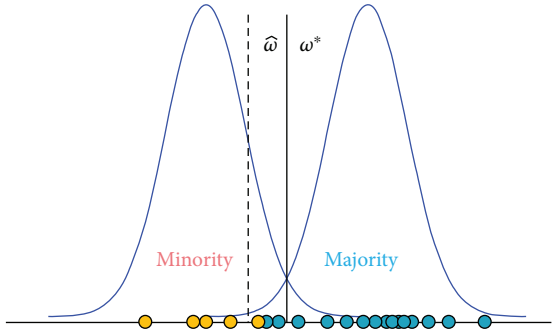


Figure 1: Bias of a linear separator.

reference point. The synthetic instance is generated by first multiplying the difference between the feature vector of the selected neighbor and x_i with a random value within the range [0,1]. Then, the following vector is added to x_i:

$$x_{\text{new}} = x_i + (\widehat{x}_i - x_i) \times \delta, \tag{1}$$

where $\widehat{x}_i$ is one of the k-nearest neighbors for x_i and $\delta \in [0, 1]$ is a random number. In the implementation of SMOTE, there are two key parameters for controlling the amount of oversampling of the minority class and undersampling of the majority classes, that is, α and γ. For each case belonging to the minority class in the original dataset, $\alpha/100$ new minority samples will be generated. The parameter γ controls the proportion of cases of the majority class that will be randomly selected for the final "balanced" dataset. This proportion is calculated with respect to the number of newly generated minority class cases. One potential drawback of SMOTE is that it generates the same number of synthetic data samples for each original minority example without considering neighboring examples, which may increase the occurrence of overlaps between classes [26].

2.2.2. Algorithm-Level Approaches. Instead of rebalancing the class distribution at the data level, some solutions have been based on biasing the existing classifiers at the algorithm level. One popular approach is to use a cost-sensitive learning method [27, 28], which considers the costs of misclassified instances and minimizes the total misclassification cost.

(1) Cost-Sensitive Learning. Unlike the rebalancing strategy, cost-sensitive learning does not directly create a balanced class distribution. Instead, it highlights the imbalanced learning problem using a cost matrix that describes the cost of misclassification in a particular scenario. Cost-sensitive learning is motivated by the observation that most real applications do not have a unified cost for misclassification [28]; therefore, the cost matrix needs to be determined beforehand, which is a major limitation. In other words, this method evaluates the cost associated with misclassifying observations [29–32]. An illustration of a cost matrix can be found in Table 1, where C(FN) and C(FP) correspond to the costs associated with a false negative (FN) and a false positive (FP), respectively. Specifically, $\text{C(FN)} > \text{C(FP)}$ defines an imbalanced classification.

Table 1: Cost matrix.

		Prediction	
		Positive	Negative
Actual	Positive	0	C(FN)
	Negative	C(FP)	0

The goal of cost-sensitive learning is to choose the classifier with the lowest total cost, that is,

$$\text{Total cost} = \text{C(FN)} \times \text{FN} + \text{C(FP)} \times \text{FP}. \tag{2}$$

It is worth noting that if a parametric model-based classifier (e.g., logistic regression) is applied, then choosing a classifier with the lowest total cost can be done by varying the threshold based on the loss function (cost of false negatives to false positives) in the training data; this is the empirical thresholding method [33]. For example, in a logistic regression, let $\mathbf{x} \in \mathbb{R}^d$ denote the d-dimensional vector of explanatory variables and let $y \in Y$ be the corresponding binary response (1 for minority and 0 for majority). The basic form of the posterior probability estimated via a linear function in $\mathbf{x}$ is as follows:

$$\begin{aligned} \Pr(y = 1 \mid \mathbf{x} = x) &= \frac{\exp\left(\beta_0 + \beta^T x\right)}{1 + \exp\left(\beta_0 + \beta^T x\right)}; \\ \Pr(y = 0 \mid \mathbf{x} = x) &= \frac{1}{1 + \exp\left(\beta_0 + \beta^T x\right)}. \end{aligned} \tag{3}$$

Given the monotone transformation, we have the following:

$$\log \frac{\Pr(y = 1 \mid \mathbf{x} = x)}{\Pr(y = 0 \mid \mathbf{x} = x)} = \beta_0 + \beta^T x. \tag{4}$$

The predicted log ratio is a hyperplane defined by $\{x \mid \widehat{\beta}_0 + \widehat{\beta}^T x = \theta\}$, where θ is the tuning threshold for choosing a classifier with the lowest total cost.

All these methods deal with imbalanced classification by directly or indirectly rebalancing the class distribution in a dataset. Various studies have presented (sometimes conflicting) viewpoints on the usefulness of different rebalancing strategies, and a comprehensive discussion can be found in [20, 26, 34].

2.3. Framework for Learning Data with Imbalanced Distribution Using Rebalancing Strategies. In this study, we develop a framework for learning healthcare data with imbalanced distribution by incorporating different rebalancing strategies and offering guideline procedure. Figure 2 shows the framework for the entire learning procedure. This framework consists of two stages: selecting a base classifier and using the base classifier to implement rebalancing strategies. In the first stage, a base classifier is selected from a set of candidates by evaluating each classifier's performance metrics (e.g., recall can be used to evaluate whether the training classifier has correctly classified minority instances). The

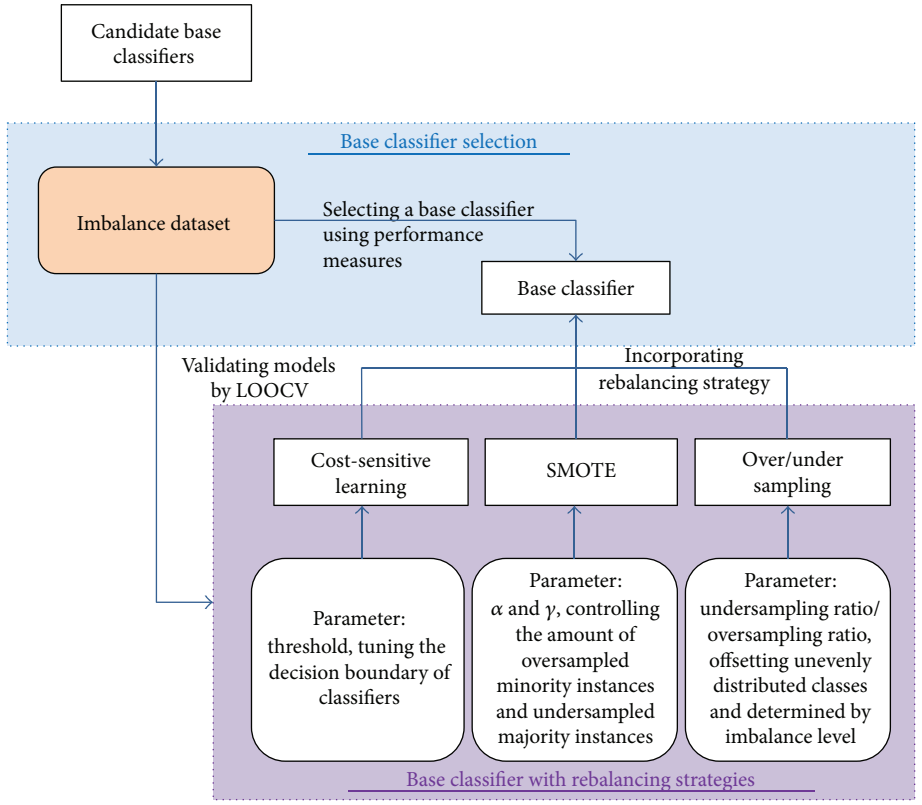


Figure 2: Framework of the learning procedure.

candidate classifiers can be either linear or nonlinear methods for binary classification (e.g., logistic regression (LR), support vector machine, decision tree, and linear discriminant analysis). In the second stage, rebalancing strategies (including resampling and cost-sensitive learning methods) can be combined with the base classifier and implemented across a range of parameters. The possible parameter sets must be designed based on the guidelines for designing the parameter/threshold of each rebalancing strategy (e.g., the oversampling ratio for an oversampling strategy should be determined based on the imbalance level). Leave-one-out cross-validation (LOOCV), which is a fair way to properly estimate model prediction performance, is used to evaluate the performance of classifiers in these stages.

3. Materials and Methods

We carry out a set of experiments to verify the effectiveness of our developed framework using a practical healthcare imbalanced dataset. We present the selected imbalanced binary classification problem, our experimental design, and the evaluation criteria used in this study as follows.

3.1. Case Description: LASA Cases. Medication names that look alike and sound alike have been recognized as the most common cause of medication errors [3]. Furthermore, 1.4% of errors due to LASA drug mix-ups have resulted in adverse and harmful patient outcomes [35, 36]. The timely and accurate identification of medication errors due to LASA drug mix-ups would reduce the medical risk to patients. Wong (2016) [5] used GPSA medical incident reports to construct classifiers for detecting LASA cases and acknowledged the challenges arising from the imbalanced classification of patient safety incident data. In our experiments, we evaluate our proposed imbalanced classification framework for detecting LASA cases using Wong's dataset [5]. The raw dataset is unstructured, as the medical incident reports are in free text format. The structured dataset used in the subsequent modeling and evaluation is a 227×8 dataset with 48 minority cases and 179 majority cases [5]. We thus regard LASA and non-LASA cases as minority and majority classes, respectively.

3.2. Base Classifiers. We compare the performance of several conventional classifiers, including the logistic regression (LR) [37], support vector machine with linear kernel (L.SVM), support vector machine with radial kernels (R.SVM) [38], and decision tree (DT) [39]. Many healthcare applications have used these classifiers due to their simplicity, interpretability, and computation efficiency [4, 5, 40]. In this study, we directly apply and validate these methods on the dataset and select the most effective classifier for detecting LASA cases as the base classifier. This base classifier is then combined with several rebalancing strategies in the second stage of the study.

3.3. Experiment Design. We investigate the performance of classifiers under various rebalancing strategies, including oversampling, undersampling, SMOTE, and cost-sensitive learning. As parameter settings usually have a significant

impact on a classifier's performance, they are thoroughly assessed in the implementation. 1) For oversampling, the oversampling ratio (number of oversampled minority instances/Number of minority instances) is set within the interval [1, 5] (*ratio1*). The interval bounds are set based on the degree to which the dataset is imbalanced ($179/48 \approx 3.7$) to reach a relatively balanced class distribution. 2) For undersampling, the undersampling ratio (number of undersampled majority instances/Number of majority instances) is set within the interval [0,1] (*ratio2*). Similarly, the interval bounds are set based on the degree to which the dataset is imbalanced. 3) For SMOTE, there are two related parameters, α and γ, which control the amount of oversampled minority instances and undersampled majority instances, respectively. 4) For cost-sensitive learning, the parameter *threshold* tunes the decision boundary of the classifiers. The detailed parameter settings are shown in Table 2.

We use LOOCV to evaluate how well the classifiers detect LASA cases, as the results of LOOCV are reproducible [41]. Five hundred replications are carried out for each set; the justification is given in the appendix. All the experiments are implemented in the R v3.3.1 (64-bit) platform using the "MASS," "e1071," "cvTools," "plyr," "DMwR," and "tree" packages [42].

3.4. Performance Evaluation Criteria. Appropriate evaluation criteria are crucial for assessing the binary classification performance of the methods. Common evaluation criteria include accuracy, recall, precision, specificity, and so on As the minority class may bias the decision boundary and has little impact on accuracy [40], we focus on performance evaluation metrics recall, precision, *F*-score, and specificity. The confusion matrix is given in Table 3.

$$\begin{aligned} \text{Precision} &= \frac{\text{TP}}{\text{TP} + \text{FP}}, \\ \text{Recall} &= \frac{\text{TP}}{\text{TP} + \text{FN}}, \\ F\text{-score} &= 2 \cdot \frac{\text{precision recall}}{\text{precision} + \text{recall}}, \\ \text{Specificity} &= \frac{\text{TN}}{\text{FN} + \text{TN}}. \end{aligned} \tag{5}$$

In assessing information retrieval, recall denotes the percentage of retrieved objects that are relevant; in the context of imbalanced classification, that is the percentage of correctly classified minority instances. Precision denotes the percentage of relevant objects that are identified for retrieval. *F*-score represents a harmonic mean between recall and precision. Specificity denotes the percentage of correctly classified majority instances. In many detection tasks, recall is the primary measure, as identifying rare but significant cases in massive unstructured healthcare datasets is our major concern. As there is always a tradeoff between recall and specificity, indiscriminately improving recall can result in a significant amount of false alarms, reflected in low specificity scores and poor overall classification accuracy. Therefore, in this study, the overall classification accuracy is controlled by specifying accuracy above 80% to avoid bias when applying rebalancing strategies.

Table 2: Parameter settings for algorithms

(a)

A.1 Random oversampling + LR								
Settings for *ratio1*								
ratio1	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0
A.2 Random undersampling + LR								
Settings for *ratio2*								
ratio2	1/1.5	1/2.0	1/2.5	1/3.0	1/3.5	1/4.0	1/4.5	1/5.0

(b)

A.3 SMOTE + LR							
Settings for α and γ							
α	400	300	200	100	200	100	100
γ	100	100	100	100	200	200	300

(c)

A.4 Cost-sensitive learning + LR							
Settings for *threshold*							
threshold	0.5	0.25	0	−0.25	−0.5	−1	−1.5	−2

Table 3: Confusion matrix.

	Condition positive	Condition negative
Test outcome positive	True positive (TP)	False positive (FP)
Test outcome negative	False negative (FN)	True negative (TN)

4. Results

4.1. Selection of Base Classifier. As shown in Table 4, all the conventional classifiers achieve good overall classification accuracy (above 80%). Among these classifiers, LR performs best in detecting LASA cases (recall = 0.521), which are our cases of interest. LR is also superior to other classifiers in terms of the synthesized measure ($F - \text{score} = 0.595$). These results are consistent with the conclusion in [5]. We thus select LR as the base classifier. However, it should be noted that the capability of LR for detecting LRSA cases is still unsatisfactory, due to the challenges associated with an imbalanced dataset.

4.2. Performance of Classifiers with Different Rebalancing Strategies. In this subsection, we investigate the effectiveness of the proposed rebalancing strategies. As described in the previous subsection, we adopt LR as the base classifier.

4.2.1. Experiment Results: Data-Level Approach. We examine the effectiveness of combining LR with various rebalancing strategies. Figure 3 compares the effectiveness of different data-level approaches for detecting LASA cases. As LASA

Table 4: Performance of conventional classifiers.

	LR	L.SVM	DT	R.SVM
Recall	0.521	0.479	0.375	0.396
Precision	0.694	0.767	0.750	0.792
F-score	0.595	0.590	0.500	0.528
Accuracy	0.850	0.859	0.841	0.850

cases are the minority class in the target dataset, the recall value indicates the method's accuracy. As shown in Figure 3, recall increases as the oversampling or undersampling ratio increases. In other words, the accuracy in the detection of LASA cases can be improved by making the class distribution more balanced, either by enlarging the size of the minority class or reducing the size of the majority class. However, there is always a tradeoff between recall and specificity/precision; the specificity/precision decreases as the recall grows. Compared with LR alone, LR in conjunction with resampling strategies gives a superior performance.

We then compare the classifiers with the best performance, all of which achieve an overall classification accuracy within the interval [0.82, 0.85], as shown in Table 5. Specifically, LR combined with oversampling improves the recall and *F*-score by 40.50 and 8.40%, respectively, under the setting $ratio1 = 3.5$; LR combined with undersampling improves the recall by 6.53% and decreases the *F*-score by 3.36%, respectively, under the setting $ratio2 = 1/1.5$; and LR combined with SMOTE outperforms all of the classifiers, improving recall and *F*-score by 45.30% and 11.76%, respectively, under the settings $\alpha = 200$ and $\gamma = 100$. As can be seen from the results, recall can be significantly improved with only a slight sacrifice (around 1.5%) in overall classification accuracy.

To evaluate the robustness of the three approaches, we plot their receiver-operating characteristic (ROC) curves, that is, true positive rate (sensitivity) against false positive rate (1 − specificity), as shown in Figure 4. The blue dashed line in each plot describes the performance of a "completely random guess" for the class of observation (i.e., the no-discrimination line from coordinates (0,0) to (1,1)), and the red line describes the ROC curve of the base classifier (pure logistic regression) for comparison purposes. A good classification method should yield points in the upper region or near the coordinate (0,1). As shown in Figure 4, all of the plots of the true positive rate against the false positive rate are above the no-discrimination line, indicating that LR combined with resampling strategies effectively reduces the effects of the two-class classification problem. The ROC curve of LR combined with SMOTE is closer to coordinate (0,1) than the ROC curve of the pure LR, indicating its superior ability to detect LASA cases.

4.2.2. Experiment Results: Algorithm-Level Approach. We also apply the cost-sensitive learning method to the detection of LASA cases. Again, LR is used as a base classifier. Figure 5 shows the classification results under various parameter settings. A smaller threshold value indicates that the decision boundary is more biased toward the majority class, that is, the non-LASA class, which increases the probability that an unknown case will be identified as a LASA case. As can be seen from Figure 5, as the threshold approaches the majority class, the recall increases and the precision decreases. The algorithm achieves the best performance when the threshold is −1; at this level, the recall and *F*-score are 14.21% and 2.93% higher, respectively, than when only the base classifier is used.

5. Discussion

5.1. Key Findings. In this study, we develop a framework for analyzing imbalanced data using rebalancing strategies. We test the effectiveness of various rebalancing strategies on a medical incident reports dataset. We conduct a comparative analysis of techniques for automatically detecting LASA cases (an imbalanced classification problem) using classifiers combined with different rebalancing strategies, including both data- and algorithm-level approaches. As there is always a tradeoff between recall and specificity, indiscriminately improving recall can result in a significant number of false alarms, reflected in low specificity and poor overall classification accuracy. The methods developed in this study maintain the overall classification accuracy at an acceptable level (acc uracy > 80%) by applying rebalancing strategies.

The results show that data-level approaches, including oversampling, undersampling, and SMOTE, are better for detecting LASA cases than algorithm-level approaches, perhaps due to the uncertainty and inconsistency of the cost matrix in training and testing the dataset. Among the data-level approaches, combining the base classifier with SMOTE, achieves the best performance; it improves the detection accuracy of LASA cases by 43.2% compared with the base classifier alone, without much loss of overall classification accuracy. There are two explanations for this result. (1) As discussed in [11], an unavoidable consequence of undersampling is a loss of information. As our dataset contains only 227 cases, randomly undersampling the majority class can result in incomplete information, which affects decision boundary learning. This explains why a random oversampling strategy generally performs better than a random undersampling strategy on small datasets. (2) Oversampling through the random replication of the minority class sample usually creates very specific rules, leading to model overfitting [7]. With replication, the decision region for the minority class can become smaller and more specific. In contrast, SMOTE builds larger decision regions that contain nearby minority class points, resulting in a higher recall. Figure 6 summarizes the changes in accuracy and recall of the different classifiers with their best performance. Taking the base classifier as a reference, it can be observed that in all the tested rebalancing strategies, the increase in recall is significantly higher than the decrease in overall detection accuracy. Specifically, the base classifier combined with SMOTE achieves the greatest increase in recall (45.3%) and the smallest decrease (1.5%) in accuracy.

It should be noted that there is no universal solution for all problems. Although SMOTE outperforms other rebalancing strategies for our dataset, it may have a higher

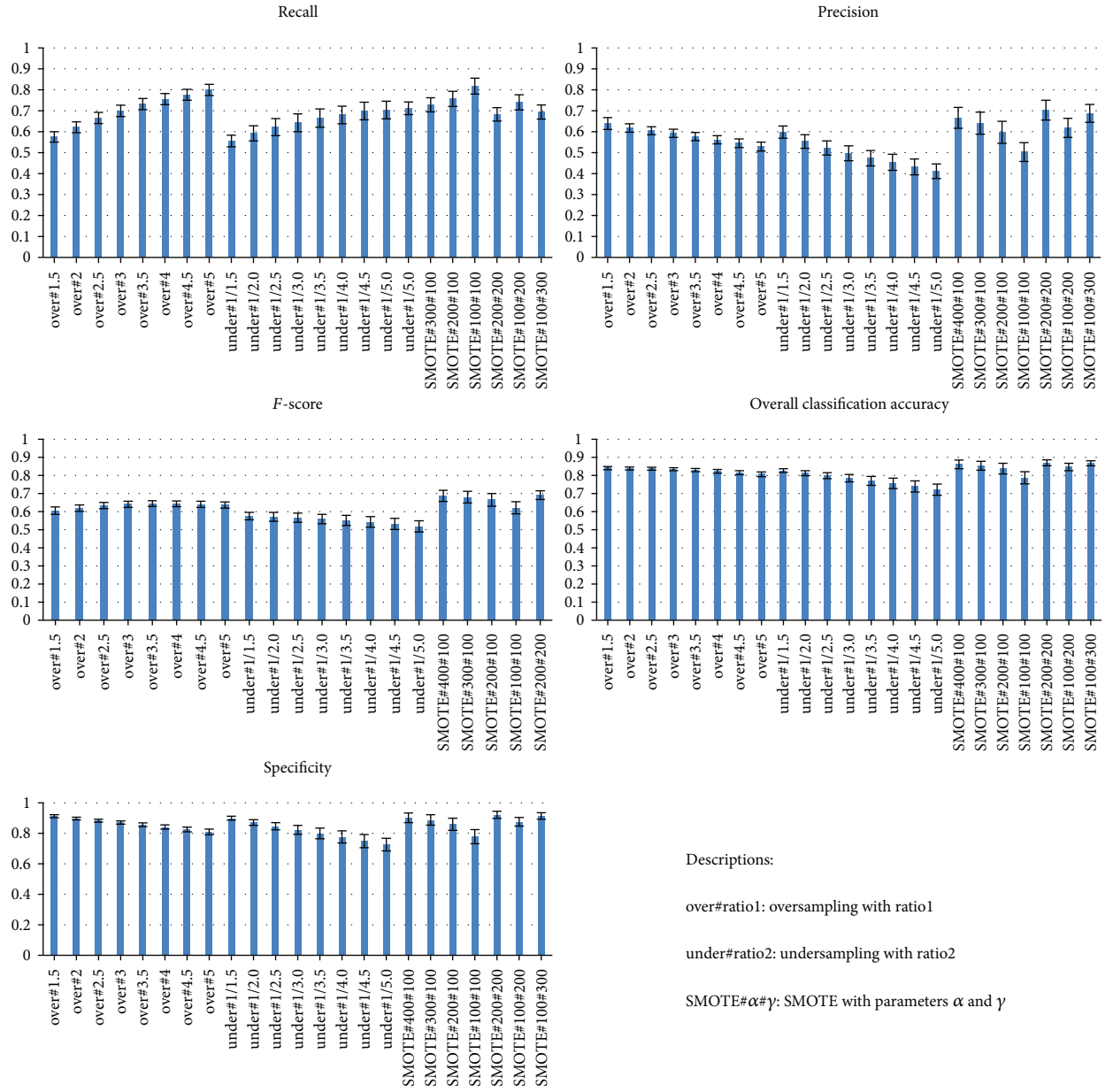
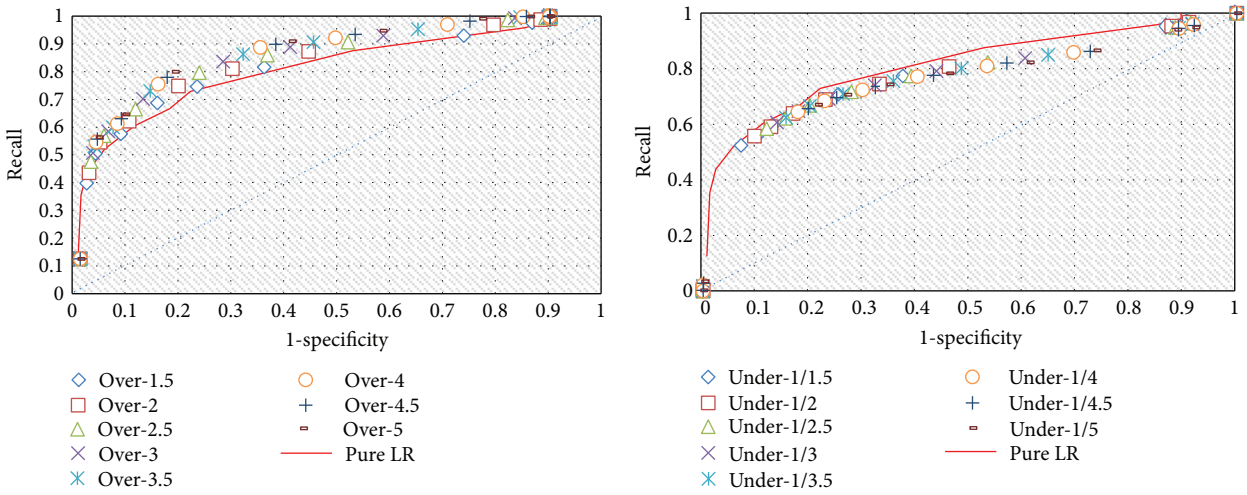


Figure 3: Comparison of data-level approaches.

Table 5: Comparison of classifiers with the best performance.

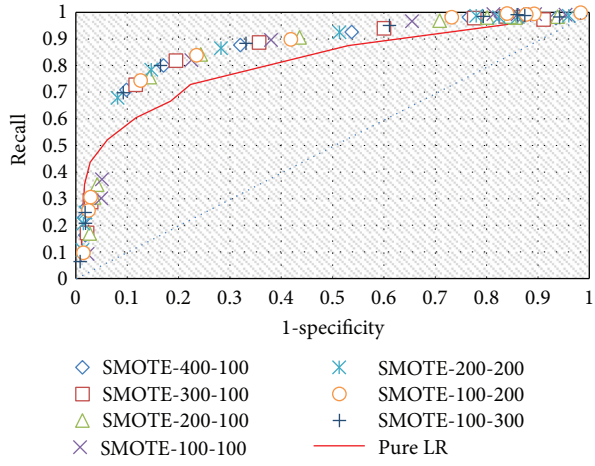
	Base classifier	Oversampling *ratio1* = 3.5 (% increase over LR)	Undersampling *ratio 2* = 1/1.5 (%increase over LR)	SMOTE $\alpha = 200$, $\gamma = 100$ (% increase over LR)
Recall	0.521	0.732 (40.50%)	0.555 (6.53%)	0.757 (45.30%)
Precision	0.694	0.577 (−16.86%)	0.598 (−13.83%)	0.597 (−13.98%)
F-score	0.595	0.645 (8.40%)	0.575 (−3.36%)	0.665 (11.76%)
Overall classification accuracy	0.850	0.829 (−2.47%)	0.826 (−2.82%)	0.837 (−1.53%)

computation cost. Due to the considerable growth in the size of the training dataset caused by the addition of synthetic samples, the training time for the resulting balanced data would be relatively higher than that for the original data. Again, as the size of our dataset is not massive, the increased computation cost is negligible. However, the overall decision-making procedure would be time costly for datasets with huge sizes and high dimensionality. In addition, algorithm-level approaches such as the cost-sensitive method may outperform data-level approaches if the cost matrices for training and testing the data are empirically known. It is important to evaluate the performance of different rebalancing strategies



(a) LR with random oversampling strategy

(b) LR with random undersampling strategy



(c) LR with SMOTE

Figure 4: ROC curves of LR with different resampling strategies.

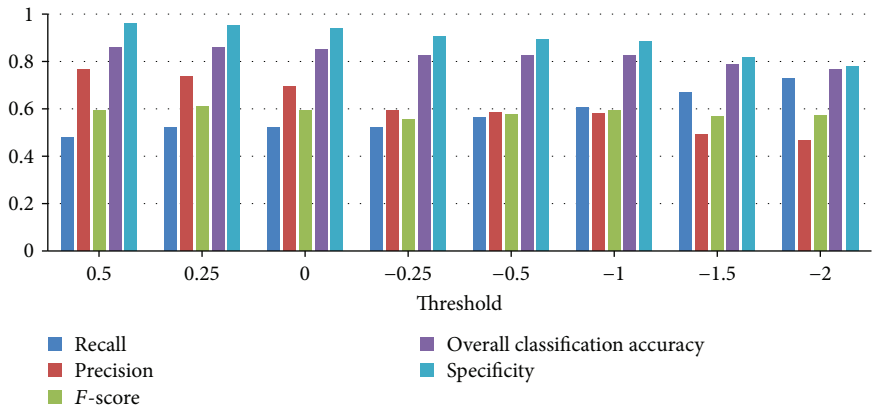


Figure 5: Comparison of cost-sensitive learning methods with various thresholds.

under the condition of no prior knowledge of data scale and cost matrices.

Nowadays, more and more crowdsourced medical data are becoming available due to the fast growth in health-related applications. As a result, techniques for analyzing big health data offer a promising and practical research direction [43]. Detecting rare but significant healthcare events through statistical data analytics may reduce treatment costs, avoid preventable diseases, and improve care quality in general. One typical application is classifying medical incident

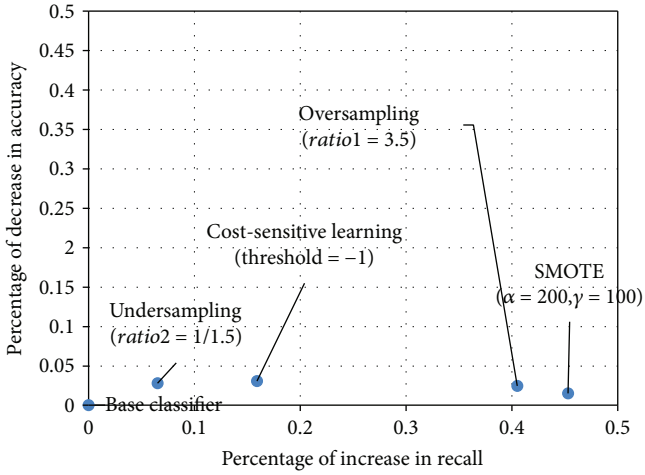


Figure 6: Summary of changes in recall and accuracy for difference classifiers.

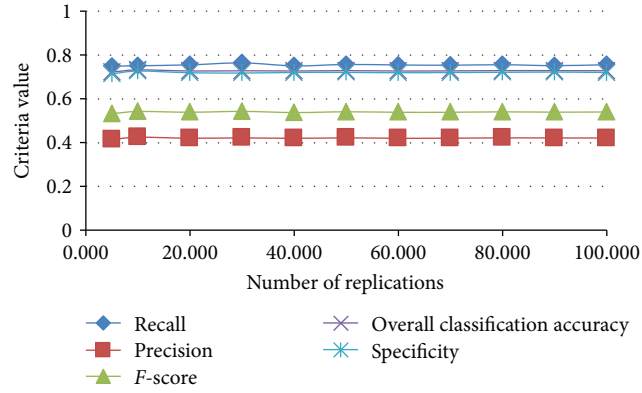


Figure 7: Performance evaluation of increasing numbers of replications.

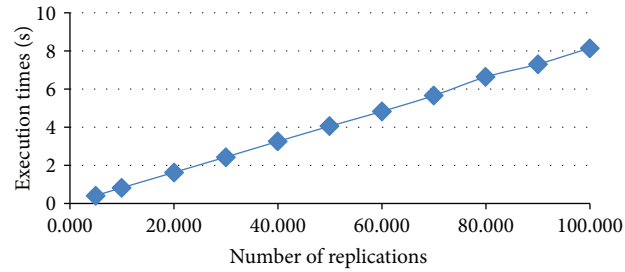


Figure 8: Execution time versus replications.

reports. Classifying incident reports at a granular level, identifying specific incident reports related to major adverse events, and discovering vital information hidden in the reports are all crucially important steps for improving patient safety. In general, the overall procedure for identifying medical incident reports involves structuring the unstructured text data using text mining to extract key terms [44, 45] and then constructing classifiers on the structured dataset for detecting specific medical incident reports. However, the imbalanced data property of incident report datasets makes such detection a challenging task.

Classifying imbalanced datasets has been recognized as a common problem in healthcare data analytics applications, such as cancer diagnostics, medical incident reports detection, and disease risk prediction. Typically, in these applications, correctly detecting minority class instances is crucial, as they correspond to high-impact events. However, the imbalance property of these datasets makes this a challenging task. Most standard classifiers, such as logistic regression and support vector machine, implicitly assume the equal occurrence of both classes, and these methods are designed for maximizing the overall classification accuracy. As a result, they favor the majority class, resulting in poor sensitivity toward the minority class. The findings in this study may help improve the detection of medical incidents caused by LASA drug mix-ups in imbalanced datasets and may eventually eliminate the need for manual identification of similar mediation-related harmful incidents. It is worth noting that the rebalancing strategies discussed in this study are not limited to medical incident report detection; they have a broad range of applications involving classification of datasets with similar imbalanced data properties.

5.2. Future Directions. In this study, we examine a two-class imbalanced classification problem using a simple illustrative example. The developed framework is potentially useful for multiclass classification problems, that is, when there are multiple classes of unequally distributed size. However, it would be difficult to learn the informed boundaries between classes in this scenario, as one has to find equilibrium between class size and detection significance when implementing rebalancing strategies. In the future, we will investigate the multiclass imbalanced classification problem in healthcare data analytics applications, such as the automatic detection of multiple types of medical incident reports. We also plan to develop an R package incorporating the imbalanced classification framework, which should considerably benefit both researchers and practitioners faced with the imbalanced classification problem in healthcare data analytics.

6. Conclusion

Detecting rare but significant healthcare events in massive unstructured datasets is now a common task in healthcare data analytics. This study is the first systematic attempt to identify rare events in unstructured healthcare datasets that have imbalanced distribution. We develop a classification framework that incorporates various rebalancing strategies for healthcare data analytics and provide some guidelines for tackling similar problems.

Appendix

Parameter Settings for Algorithms

The sampling schemes should be repeated to obtain unbiased evaluation results. To estimate the least number of sampling replications necessary (*rep*) and assess the experiment scale, we execute the algorithm "Random oversampling + LR" (*rat*

io1 = 5) with an increasing number of replications. Figure 7 displays the averaged evaluation result with an increasing number of replications. As can be seen, the result tends to be stable when the number of replications approaches 100. It is worth noting that the execution time grows linearly as the replication increases, which can be observed in Figure 8. In the implementation, the number of replications is fixed at 500 to ensure a reliable evaluation.

Abbreviations

LASA: Look-alike sound-alike
LOOCV: Leave-one-out cross-validation
SMOTE: Synthetic minority oversampling technique
FN: False negative
FP: False positive
LR: Logistic regression
L.SVM: Support vector machine with linear kernel
R.SVM: Support vector machine with radial kernel
DT: Decision tree
ROC: Receiver operating characteristic
EHRs: Electronic health records.

Data Availability

All of the datasets used in this study are publicly available at the Global Patient Safety Alerts (GPSA) system.

Ethical Approval

Ethical approval is not applicable.

Disclosure

The funding agreements ensured the authors' independence in designing the study, interpreting the data, and writing and publishing the article.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported in part by the RGC Theme-Based Research Scheme (TBRS) (no. T32-102/14-N), the National Natural Science Foundation of China (NSFC) (no. 71420 107023), and Japan Society for the Promotion of Science KAKENHI (no. 18H03336).

References

[1] W. Raghupathi and V. Raghupathi, "Big data analytics in healthcare: promise and potential," *Health Information Science and Systems*, vol. 2, no. 1, 2014.

[2] K. L. Tsui and Y. Zhao, "Discussion of "analyzing behavioral big data: methodological, practical, ethical, and moral issues"," *Quality Engineering*, vol. 29, pp. 79–83, 2016.

[3] L. K. McCoy, "Look-alike, sound-alike drugs review: include look-alike packaging as an additional safety check," *The Joint Commission Journal on Quality and Patient Safety*, vol. 31, no. 1, pp. 47–53, 2005.

[4] M.-S. Ong, F. Magrabi, and E. Coiera, "Automated categorisation of clinical incident reports using statistical text classification," *BMJ Quality & Safety*, vol. 19, no. 6, article e55, 2010.

[5] Z. S. Y. Wong, "Statistical classification of drug incidents due to look-alike sound-alike mix-ups," *Health Informatics Journal*, vol. 22, no. 2, pp. 276–292, 2016.

[6] G. M. Weiss, "Mining with rarity: a unifying framework," *ACM SIGKDD Explorations Newsletter*, vol. 6, no. 1, pp. 7–19, 2004.

[7] R. C. Holte, L. E. Acker, and B. W. Porter, "Concept learning and the problem of small disjuncts," in *IJCAI '89 Proceedings of the 11th international joint conference on Artificial intelligence*, vol. 1, pp. 813–818, Detroit, MI, USA, August 1989.

[8] E. Sarioglu, H. A. Choi, and K. Yadav, "Clinical report classification using natural language processing and topic modeling," in *2012 11th International Conference on Machine Learning and Applications*, pp. 204–209, Boca Raton, FL, USA, December 2012.

[9] N. Emanet, H. R. Öz, N. Bayram, and D. Delen, "A comparative analysis of machine learning methods for classification type decision problems in healthcare," *Decision Analytics*, vol. 1, no. 1, p. 6, 2014.

[10] J. Bates, S. J. Fodeh, C. A. Brandt, and J. A. Womack, "Classification of radiology reports for falls in an HIV study cohort," *Journal of the American Medical Informatics Association*, vol. 23, no. e1, pp. e113–e117, 2016.

[11] Y. Tang, Y. Q. Zhang, N. V. Chawla, and S. Krasser, "SVMs modeling for highly imbalanced classification," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 39, no. 1, pp. 281–288, 2009.

[12] M. Galar, A. Fernandez, E. Barrenechea, H. Bustince, and F. Herrera, "A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 42, no. 4, pp. 463–484, 2012.

[13] H. Hassanzadeh, T. Groza, A. Nguyen, and J. Hunter, "Load balancing for imbalanced data sets: classifying scientific artefacts for evidence based medicine," in *PRICAI 2014: Trends in Artificial Intelligence. PRICAI 2014*, D. N. Pham and S. B. Park, Eds., vol. 8862 of Lecture Notes in Computer Science, pp. 972–984, Springer, Cham, 2014.

[14] E. Ortiz, G. Meyer, and H. Burstin, "Clinical informatics and patient safety at the Agency for Healthcare Research and Quality," *Journal of the American Medical Informatics Association*, vol. 9, no. 90061, Supplement 6, pp. S2–S7, 2002.

[15] M. Khalilia, S. Chakraborty, and M. Popescu, "Predicting disease risks from highly imbalanced data using random forest," *BMC Medical Informatics and Decision Making*, vol. 11, no. 1, pp. 1–13, 2011.

[16] A. Estabrooks, T. Jo, and N. Japkowicz, "A multiple resampling method for learning from imbalanced data sets," *Computational Intelligence*, vol. 20, no. 1, pp. 18–36, 2004.

[17] C. Drummond and R. C. Holte, "C4.5, class imbalance, and cost sensitivity: why under-sampling beats over-sampling," in *Proceedings of the ICML'03 Workshop on Learning from Imbalanced Data Sets*, pp. 1–8, Washington, DC, USA, 2003.

[18] J. Zhang and I. Mani, "KNN approach to unbalanced data distributions: a case study involving information extraction," in

Proceedings of the ICML'03 Workshop on Learning from Imbalanced Data Sets, Washington, DC, USA, 2003.

[19] A. H. S. Solberg and R. Solberg, "A large-scale evaluation of features for automatic detection of oil spills in ERS SAR images," in *Geoscience and Remote Sensing Symposium, 1996. IGARSS '96. 'Remote Sensing for a Sustainable Future.', International*, pp. 1484–1486, Lincoln, NE, USA, May 1996.

[20] N. Japkowicz, "The class imbalance problem: significance and strategies," in *Proceedings of the 2000 International Conference on Artificial Intelligence: Special Track on Inductive Learning*, pp. 111–117, Las Vegas, NV, USA, 2000.

[21] M. Kubat and S. Matwin, "Addressing the curse of imbalanced training sets: one-sided selection," in *Proceedings of the 14th International Conference on Machine Learning, 1997*, pp. 79–186, Nashville, TN, USA, 1997.

[22] T. Jo and N. Japkowicz, "Class imbalances versus small disjuncts," *ACM SIGKDD Explorations Newsletter*, vol. 6, no. 1, pp. 40–49, 2004.

[23] G. E. A. P. A. Batista, R. C. Prati, and M. C. Monard, "A study of the behavior of several methods for balancing machine learning training data," *ACM SIGKDD Explorations Newsletter*, vol. 6, no. 1, pp. 20–29, 2004.

[24] C. Phua, D. Alahakoon, and V. Lee, "Minority report in fraud detection: classification of skewed data," *ACM SIGKDD Explorations Newsletter*, vol. 6, no. 1, pp. 50–59, 2004.

[25] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, no. 1, pp. 321–357, 2002.

[26] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, no. 9, pp. 1263–1284, 2009.

[27] C. Elkan, "The foundations of cost-sensitive learning," in *IJCAI'01 Proceedings of the 17th international joint conference on Artificial intelligence*, pp. 973–978, Seattle, WA, USA, August 2001.

[28] P. Domingos, "MetaCost: a general method for making classifiers cost-sensitive," in *KDD '99 Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 155–164, San Diego, CA, USA, August 1999.

[29] S. Zhang, "Cost-sensitive classification with respect to waiting cost," *Knowledge-Based Systems*, vol. 23, no. 5, pp. 369–378, 2010.

[30] S. Zhang, "Decision tree classifiers sensitive to heterogeneous costs," *Journal of Systems and Software*, vol. 85, no. 4, pp. 771–779, 2012.

[31] T. Wang, Z. Qin, S. Zhang, and C. Zhang, "Cost-sensitive classification with inadequate labeled data," *Information Systems*, vol. 37, no. 5, pp. 508–516, 2012.

[32] T. Wang, Z. Qin, Z. Jin, and S. Zhang, "Handling over-fitting in test cost-sensitive decision tree learning by feature selection, smoothing and pruning," *Journal of Systems and Software*, vol. 83, no. 7, pp. 1137–1147, 2010.

[33] C. X. Ling and V. S. Sheng, "Cost-sensitive learning and the class imbalance problem," in *Encyclopedia of Machine Learning*, Springer, Berlin, 2008.

[34] N. V. Chawla, "Data mining for imbalanced datasets: an overview," in *Data Mining and Knowledge Discovery Handbook*, O. Maimon and L. Rokach, Eds., pp. 853–867, Springer, Boston, MA, USA, 2005.

[35] S. Kim, "Pharmacopeia 8th Annual MEDMARX(R) report indicates look-alike/sound-alike drugs," 2008, August 2016, http://www.reuters.com/article/2008/01/29/idUS210935+29-Jan-2008+PRN20080129.

[36] M. Kaufman, "Preventable medication errors: look-alike/sound-alike mix-ups," 2011, August 2016, http://formularyjournal.modernmedicine.com/formulary/article/articleDetail.jsp?id=579387.

[37] D. R. Cox, "The regression analysis of binary sequences," *Journal of the Royal Statistical Society Series B*, vol. 20, pp. 215–242, 1958.

[38] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.

[39] L. Breiman, J. Friedman, R. Olshen, and C. Stone, *Classification and Regression Trees*, Wadsworth and Brooks, Belmonte, CA, USA, 1984.

[40] Y. Sun, A. K. C. Wong, and M. S. Kamel, "Classification of imbalanced data: a review," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 23, no. 4, pp. 687–719, 2009.

[41] G. Seni and J. Elder, "Ensemble methods in data mining: improving accuracy through combining predictions," *Synthesis Lectures on Data Mining and Knowledge Discovery*, vol. 2, no. 1, pp. 1–126, 2010.

[42] R Development Core Team, R, *A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2013.

[43] S. Schneeweiss, "Learning from big health care data," *The New England Journal of Medicine*, vol. 370, no. 23, pp. 2161–2163, 2014.

[44] I. Feinerer and K. Hornik, *tm: Text Mining Package*, 2015, https://cran.r-project.org/package=tm.

[45] I. Feinerer, K. Hornik, and D. Meyer, "Text mining infrastructure in R," *Journal of Statistical Software*, vol. 25, no. 5, 2008.