Why Does Rebalancing Class-Unbalanced Data Improve AUC for Linear Discriminant Analysis?

Jing-Hao Xue and Peter Hall

Abstract—Many established classifiers fail to identify the minority class when it is much smaller than the majority class. To tackle this problem, researchers often first rebalance the class sizes in the training dataset, through oversampling the minority class or undersampling the majority class, and then use the rebalanced data to train the classifiers. This leads to interesting empirical patterns. In particular, using the rebalanced training data can often improve the area under the receiver operating characteristic curve (AUC) for the original, unbalanced test data. The AUC is a widely-used quantitative measure of classification performance, but the property that it increases with rebalancing has, as yet, no theoretical explanation. In this note, using Gaussian-based linear discriminant analysis (LDA) as the classifier, we demonstrate that, at least for LDA, there is an intrinsic, positive relationship between the rebalancing of class sizes and the improvement of AUC. We show that the largest improvement of AUC is achieved, asymptotically, when the two classes are fully rebalanced to be of equal sizes.

Index Terms—AUC, class imbalance, class rebalancing, linear discriminant analysis, oversampling, ROC, undersampling

1 Introduction

In many practical contexts, instances have to be classified into two classes of remarkably distinct sizes [1], [2], [3], for example, a minority class of fraudsters or cancer patients versus a majority class of honest customers or normal people.

In such cases, many established classifiers often trivially classify instances into the majority class, achieving an optimal overall misclassification error rate. This, however, leads to poor performance in classifying the minority class, the correct identification of which is usually of more practical interest.

To tackle this problem, researchers often first rebalance the class sizes in the training dataset, through oversampling the minority class or undersampling the majority class, and then use the rebalanced data to train the classifiers. (See the recent, comprehensive survey of rebalancing methods [3].)

Such a rebalancing strategy results in interesting empirical patterns. Using the rebalanced training data can often increase (i.e. improve) the area (AUC) under the receiver operating characteristic (ROC) curve for the original, unbalanced test data [4], [5], [6], [7]. The AUC is a widely-used quantitative measure of classification performance, but the empirical property that rebalancing increases AUC lacks theoretical justification.

Therefore, in this note, using Gaussian-based linear discriminant analysis (LDA) as the classifier, we shall show theoretically that, at least for LDA, there is an intrinsic, positive relationship between rebalancing class sizes and improving AUC. In particular, we shall demonstrate that the largest improvement in AUC can be achieved, asymptotically, when the two classes are fully rebalanced to be of equal sizes.

2 A Motivating Example

2.1 Notation

In the training dataset there are $n = n_0 + n_1$ instances with d features each, including $\{\mathbf{x}_{0i}\}_{i=1}^{n_0}$ arising from a majority (or negative) class $\mathcal{C}_0$ and $\{\mathbf{x}_{1i}\}_{i=1}^{n_1}$ from a minority (or positive) class $\mathcal{C}_1$, with $n_0 \gg n_1$.

The two classes are often assumed to have respective d-dimensional Gaussian distributions: for any training or test instance $\mathbf{x}_{jk}$ in $\mathcal{C}_j$, $\mathbf{x}_{jk} \sim \mathrm{N}(\mu_j, \Sigma_j)$ for $j = 0, 1$, where μ_j and Σ_j denote the population mean vector and covariance matrix, respectively, of class $\mathcal{C}_j$. The prevalence rate (i.e. prior probability) of $\mathcal{C}_j$ is denoted by π_j, with $\pi_0 + \pi_1 = 1$.

Typically μ_j and Σ_j are estimated by the sample mean vector $\hat{\mu}_j$ and sample within-class covariance matrix $\hat{\Sigma}_j$, and π_j is estimated by

$$\hat{\pi}_j = n_j/n \ . \tag{1}$$

2.2 An Example Showing that Rebalancing Can Improve AUC for LDA

Fig. 1 shows a motivating example, using a scatter plot and a panel of nine boxplots of AUC to illustrate visually the fact that rebalancing the training data can often improve the performance of LDA in terms of AUC [5], [6]. This example is extracted from an experiment on simulated data arising from two four-dimensional, Gaussian-distributed classes $\mathcal{C}_0$ and $\mathcal{C}_1$. With a slightly different setting, the experiment explores more rebalancing scenarios than in [6]. It includes the following four steps.

Firstly, a dataset was constructed by simulating $2n$ instances, as follows. Eighty percent were simulated from a majority class $\mathcal{C}_0$ with $\mathbf{x}_{0i} \sim \mathrm{N}(\mu_0, \Sigma_0)$ where $\mu_0 = \mathbf{0}$ and Σ_0 was the identity matrix $\mathbf{I}$, and the other 20 percent were simulated from a minority class $\mathcal{C}_1$ with $\mathbf{x}_{1i} \sim \mathrm{N}(\mu_1, \Sigma_1)$, where $\mu_1 = (-1.5, -0.75, 0.75, 1.5)^T$ and Σ_1 was a diagonal matrix with four unequal components, respectively equal to 0.25, 0.75, 1.25 and 1.75. This setting preserves generality for Gaussian-distributed classes, since we can use a linear transformation to simplify Σ_0 to $\mathbf{I}$ and diagonalise Σ_1; see [8], [9].

Secondly, the dataset was randomly split into a training set and a test set, each containing half of the instances (i.e. n instances) and maintaining the prevalence rate of each class. Such a split was repeated 100 times, and in this way 100 pairs of training and test sets were obtained.

Thirdly, each training set, which is original and unbalanced with $\hat{\pi}_1 = 20\%$, was rebalanced to predetermined proportions 30, 40, 50 or 60 percent, respectively. Such rebalancing was accomplished through random oversampling or random undersampling [3], undertaken with replacement in each case.

Finally, LDA, implemented by a function `lda` from an R package *'MASS'*, was applied to the original data and the rebalanced data, and corresponding AUCs were recorded. Each AUC was calculated via the Wilcoxon rank-sum statistic, which was implemented by a function `wilcox.test` in the R package *'stats'*.

From Fig. 1 we observe not only that rebalancing class sizes improves AUC, but also that the largest improvement is attained at around full rebalance, where the two rebalanced classes possess equal numbers of instances in the training set. The next section explores this issue theoretically.

3 Theoretical Analysis

3.1 LDA

LDA, or more precisely 'plug-in' Gaussian-based LDA, assumes that Σ_0 and Σ_1 are both equal to Σ, say, and uses a linear rule to classify a new $\mathbf{x}$ into $\mathcal{C}_1$ if $\hat{\beta}^T \mathbf{x} \geq -\hat{\beta}_0$, where

$$\hat{\beta} = \hat{\Sigma}^{-1}(\hat{\mu}_1 - \hat{\mu}_0), \tag{2}$$

- J.-H. Xue is with the Department of Statistical Science, University College London, London, WC1E 6BT, United Kingdom. E-mail: jinghao.xue@ucl.ac.uk.
- P. Hall is with the Department of Mathematics and Statistics, the University of Melbourne, VIC 3010, Australia. E-mail: halpstat@ms.unimelb.edu.au.

Manuscript received 20 June 2012; revised 9 Feb. 2014; accepted 4 Sept. 2014. Date of publication 21 Sept. 2014; date of current version 4 Apr. 2015.
Recommended for acceptance by E.P. Xing.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TPAMI.2014.2359660

This work is licensed under a Creative Commons Attribution 3.0 License. For more information, see http://creativecommons.org/licenses/by/3.0/

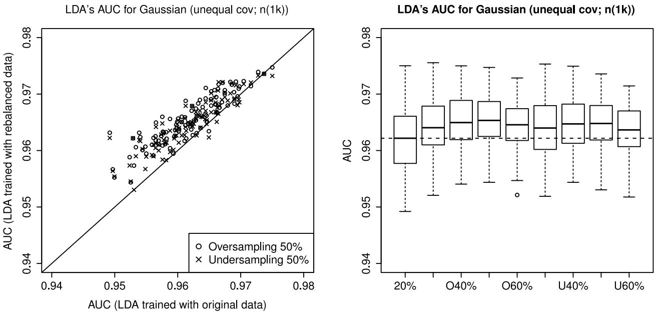

Fig. 1. AUC for data (with $n = 1{,}000$) arising from two Gaussian classes with unequal covariance matrices. Left: Scatter plot of AUC for LDA trained with fully rebalanced data (i.e. $\hat{\pi}_1 = 50\%$) vs. AUC for LDA trained with original unbalanced data (with $\hat{\pi}_1 = 20\%$). Right: Boxplots of AUC for LDA trained with the original unbalanced data and rebalanced data, in which, from left to right, the leftmost boxplot corresponds to the original data, the next four boxplots to the rebalancing scenarios 'O30'-'O60 percent' and the rightmost four boxplots to 'U30'-'U60 percent', where 'O50 percent' ('U50 percent') denotes class rebalancing through random oversampling (undersampling) the minority (majority) class such that $\hat{\pi}_1 = 50\%$, for example. The dashed horizontal line across the boxplots indicates the median AUC for the original data.

$$\hat{\beta}_0 = \log \frac{\hat{\pi}_1}{\hat{\pi}_0} - \frac{1}{2} (\hat{\mu}_1 + \hat{\mu}_0)^T \hat{\Sigma}^{-1} (\hat{\mu}_1 - \hat{\mu}_0) \ , \tag{3}$$

and $\hat{\Sigma} = \hat{\pi}_0 \hat{\Sigma}_0 + \hat{\pi}_1 \hat{\Sigma}_1$; and classifies $\mathbf{x}$ into $\mathcal{C}_0$ otherwise. We have used "hat" notation for $\hat{\beta}$ and $\hat{\beta}_0$ to indicate that they depend on the data.

Fisher's linear discriminant, often termed Fisher's LDA, is almost the same as Gaussian-based LDA, without assuming Gaussian distributions for $\mathcal{C}_j$ or equal covariance matrices. It classifies $\mathbf{x}$ into $\mathcal{C}_1$ if $\beta^T \mathbf{x} \geq c$, where β maximises the ratio

$$R_{\mathrm{LDA}} = \frac{\{\beta^T(\hat{\mu}_1 - \hat{\mu}_0)\}^2}{\beta^T \hat{\Sigma} \beta} \ ,$$

and into $\mathcal{C}_0$ otherwise, for an appropriate vector β. By solving a generalised eigenvalue problem it can be shown that, up to a constant of proportionality, β equals its counterpart in (2), and in particular equals $\alpha \hat{\Sigma}^{-1}(\hat{\mu}_1 - \hat{\mu}_0)$ where α is an arbitrary scalar. Hence, the threshold c can equal $-\alpha \hat{\beta}_0$ if the assumptions underlying Gaussian-based LDA hold, and can take another value otherwise.

3.2 Why Rebalancing Can Improve AUC for LDA

For two-class discrimination the ROC curve is a graph of the true positive rate versus the false positive rate at varying discriminant thresholds. Hence a higher AUC generally indicates a superior classifier. The AUC is equivalent to the Mann-Whitney U statistic (also termed the Wilcoxon rank-sum statistic), the Gini coefficient, Harrell's c (for concordance) index, and the probability of a correct ranking of a randomly chosen pair of positive and negative instances [10], [11], [12].

It has been shown in [13], [14] that, if

$$\mathbf{x}_{jk} \sim \mathrm{N}(\mu_j, \Sigma_j) \tag{4}$$

for the test data, then AUC (i.e. binormal AUC) can be computed as

$$\mathrm{AUC} = \Phi(\sqrt{R_{\mathrm{AUC}}}) = \Phi\left(\sqrt{\frac{\{\beta^T(\mu_1 - \mu_0)\}^2}{\beta^T(\Sigma_0 + \Sigma_1)\beta}}\right) \ , \tag{5}$$

where Φ is the cumulative distribution function of the $\mathrm{N}(0,1)$ distribution. Both Φ and the square root transformation are strictly increasing functions, and so maximising AUC is equivalent to maximising R_{AUC}, which is achieved by taking

$$\beta = \alpha(\Sigma_0 + \Sigma_1)^{-1}(\mu_1 - \mu_0) \ . \tag{6}$$

Therefore the largest value that AUC can take is

$$\mathrm{AUC}_{\max} = \Phi\left(\sqrt{(\mu_1 - \mu_0)^T(\Sigma_0 + \Sigma_1)^{-1}(\mu_1 - \mu_0)}\right) \ . \tag{7}$$

The results (6) and (7) were proved in [15], [16] using somewhat different approaches.

Of course, the argument in the previous paragraph was for cases where the true values of μ_j, Σ_j and β, rather than their estimators $\hat{\mu}_j$, $\hat{\Sigma}_j$ and $\hat{\beta}$, are used to construct the classifier. However, we shall show in Section 5 that, under mild assumptions, and writing $\widetilde{\mathrm{AUC}}$ for the value of area under the ROC curve when the classifier is constructed using $\hat{\mu}_j$ and $\hat{\Sigma}_j$,

$$\widetilde{\mathrm{AUC}} \to \mathrm{AUC}, \tag{8}$$

as the training sample sizes increase. Here, AUC on the right-hand side denotes the quantity at (5).

As the training sample sizes grow, $\hat{\mu}_j$ and $\hat{\Sigma}_j$ converge to their true values:

$$\hat{\mu}_j \to \mu_j \ , \quad \hat{\Sigma}_j \to \Sigma_j \ . \tag{9}$$

(The convergences in (9)-(14) are all in probability, and follow from laws of large numbers; (9) follows from consistency of the sample mean and sample covariance matrix, and (10)-(14) follow from (9).) Assume that the fraction $\hat{\pi}_j$, defined at (1), converges to a value $\pi_j > 0$ for $j = 0, 1$. Then

$$\hat{\Sigma} = \hat{\pi}_0 \hat{\Sigma}_0 + \hat{\pi}_1 \hat{\Sigma}_1 \to \pi_0 \Sigma_0 + \pi_1 \Sigma_1 \ . \tag{10}$$

However, if we rebalance the training data, in such a way that the version of $\hat{\pi}_j$ for the rebalanced training data equals $\frac{1}{2}$, then the version of $\hat{\Sigma}$ for those data (denote it by $\hat{\Sigma}^{\mathrm{rebal}}$) equals $\frac{1}{2}\hat{\Sigma}_0 + \frac{1}{2}\hat{\Sigma}_1$ rather than the quantity given by first identity in (10). In this case the limit relation in (10) alters to

$$\hat{\Sigma}^{\mathrm{rebal}} \to \tfrac{1}{2}(\Sigma_0 + \Sigma_1) \ , \tag{11}$$

and the limiting value of AUC, in (8), is $\mathrm{AUC}_{\max}$, at (7):

$$\widetilde{\mathrm{AUC}} \to \mathrm{AUC}_{\max} \ . \tag{12}$$

Result (12), which is also derived in Section 5, confirms that the value of AUC for the empirical, rebalanced classifier converges to the maximum possible AUC as the training samples diverge. This clarifies why, for LDA, rebalancing the training data has an intrinsic and positive relationship with the improvement of AUC for the original unbalanced test data, and why the largest improvement of AUC occurs approximately (indeed exactly, in the asymptotic limit) when the two classes are fully rebalanced to be of equal sizes.

4 Further Notes

4.1 When Σ_0 Equals Σ_1

When $\Sigma_0 = \Sigma_1$, the formulae for $\hat{\beta}$ given by the first identities in (13) and (14) provide asymptotically the same vector of coefficients β that leads to the largest AUC, since the arbitrary scalar α in (6) has no effect on the direction of the vector β, nor on the largest AUC in (7). In other words, when the two Gaussian classes have similar covariance matrices, rebalancing class sizes provides little by way of improvement in AUC for LDA. An illustration is offered by Fig. 2, where the setting is the same as that for Fig. 1 except that $\Sigma_1 = \Sigma_0 = \mathbf{I}$.

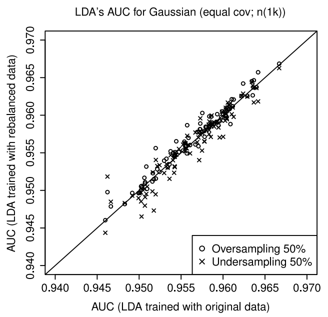

Fig. 2. AUC for data arising from two Gaussian classes with equal covariance matrices. The rest of the caption is as in Fig. 1.

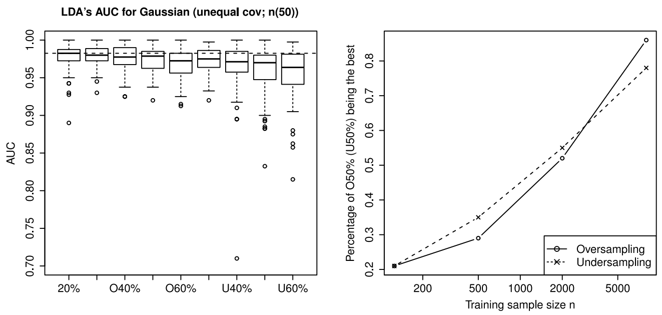

Fig. 3. Left: AUC for data with $n = 50$. The setting and rest of the caption are as in Fig. 1. Right: The percentage, for which full rebalancing (i.e. 'O50 percent' or 'U50 percent') achieves the largest improvement in our experiments, versus the training sample size n, with $n = 125, 500, 2{,}000$ and $8{,}000$.

4.2 When Data Are Univariate (i.e. $d = 1$)

When the data are univariate, the β's in (13), (14) and (6) are all scalars; these scalars lead to the same AUC. In other words, rebalancing class sizes for univariate data does not improve AUC for LDA.

4.3 When Data Are Non-Gaussian

When the Gaussian assumption $\mathbf{x}_{jk} \sim \mathrm{N}(\mu_j, \Sigma_j)$ does not hold, or more specifically when $\beta^T\mathbf{x}$ is not Gaussian, formulae (5)-(7) are no longer valid. Hence, using the β given by the first identity in (14) does not guarantee an increase in AUC relative to that for the β given by the first identity in (13). In other words, rebalancing class sizes may not improve AUC when LDA is applied to non-Gaussian data.

4.4 When the Sample Size Is Small

Our theoretical results are asymptotic, and show that, for all sufficiently large but finite datasets, there is a strict improvement in AUC resulting from rebalancing the training data. Nevertheless, when the sample size is small, these improvements cannot be guaranteed theoretically. In fact, for a small sample, *empirically* there is also no guarantee of improvement in AUC resulting from rebalancing. This is illustrated in the left-hand panel of Fig. 3, where we plot results for a randomly generated Gaussian dataset, with $n = 50$.

From Fig. 3 we can also observe the empirical result that, although full rebalancing does not necessarily maximise performance for a small sample (see the left-hand panel), it has a higher probability of achieving the largest improvement for a larger sample, compared with other proportional rebalancing schemes in our experiments (see the right-hand panel). This reflects our theoretical results.

4.5 Extensions to Other Imbalanced-Learning Techniques

Besides rebalancing the class sizes, typical imbalanced-learning techniques include strategies for adopting distinct class-dependent costs for misclassification errors, adjusting the decision threshold, and weighting training instances. Although these methods are not discussed here, we note that they are generally equivalent to class-rebalancing approaches [17], [18].

Random oversampling and undersampling are used in this note to illustrate the effect of rebalancing the training data. Numerous sampling mechanisms have been developed for rebalancing; see the seven subsections of Section 3.1 in the survey [3]. The conclusions presented here are based on (4) and (9), and hence can be extended to other sampling mechanisms that do not shift the sample mean vectors and covariance matrices (obtained using the rebalanced training data) away from their values represented by the test data.

5 Proofs of (8) and (12)

From (9), (10) and (11) and the property $\hat{\pi}_j = n_j/n \to \pi_j$ it follows that the versions $\hat{\beta}$ of LDA, respectively estimated from the unbalanced and rebalanced data, satisfy

$$\begin{aligned}\hat{\beta} &= (\hat{\pi}_0\hat{\Sigma}_0 + \hat{\pi}_1\hat{\Sigma}_1)^{-1}(\hat{\mu}_1 - \hat{\mu}_0)\\ &\to (\pi_0\Sigma_0 + \pi_1\Sigma_1)^{-1}(\mu_1 - \mu_0)\ ,\end{aligned} \tag{13}$$

$$\begin{aligned}\hat{\beta} &= \left(\hat{\Sigma}^{\mathrm{rebal}}\right)^{-1}(\hat{\mu}_1 - \hat{\mu}_0)\\ &\to 2(\Sigma_0 + \Sigma_1)^{-1}(\mu_1 - \mu_0)\ .\end{aligned} \tag{14}$$

Consider LDA when (4) holds but, instead of using the true values of μ_j and Σ_j, and a fixed β, apply LDA with perturbed values μ_j^{pert}, Σ_j^{pert} and β^{pert}, say, satisfying

$$\|\mu_j^{\mathrm{pert}} - \mu_j\| \le \epsilon\ ,\quad \|\Sigma_j^{\mathrm{pert}} - \Sigma_j\| \le \epsilon\ ,\quad \|\beta^{\mathrm{pert}} - \beta\| \le \epsilon, \tag{15}$$

for $j = 0, 1$, where we have used standard norms for vectors and matrices, and $\epsilon > 0$. Write $\mathrm{AUC}^{\mathrm{pert}}$ for the resulting value of AUC for the LDA classifier. By [13], [14] and the continuous mapping theorem, if $\beta^T(\Sigma_0 + \Sigma_1)\beta > 0$, if (15) holds, and if ϵ is small (say, $\epsilon \in (0, \epsilon_0]$), then

$$\left|\mathrm{AUC}^{\mathrm{pert}} - \mathrm{AUC}\right| \le C\epsilon, \tag{16}$$

where AUC is as at (5) and $C > 0$.

Write $\inf_\epsilon \mathrm{AUC}^{\mathrm{pert}}$ and $\sup_\epsilon \mathrm{AUC}^{\mathrm{pert}}$ for the infimum and supremum, respectively, of $\mathrm{AUC}^{\mathrm{pert}}$ over μ_j^{pert}, Σ_j^{pert} and β^{pert} such that (15) holds. Let $\mathcal{E}_\epsilon$ denote the event that (15) holds if μ_j^{pert}, Σ_j^{pert} and β^{pert} are replaced by $\hat{\mu}_j$, $\hat{\Sigma}_j$ and $\hat{\beta}$ (the latter defined by (2), with $\hat{\Sigma} = \hat{\pi}_0\hat{\Sigma}_0 + \hat{\pi}_1\hat{\Sigma}_1$ if we do not rebalance, and with $\hat{\Sigma} = \hat{\Sigma}^{\mathrm{rebal}} = \frac{1}{2}(\hat{\Sigma}_0 + \hat{\Sigma}_1)$ if we do), respectively. Put $p(\epsilon) = 1 - P(\mathcal{E}_\epsilon)$. Consider the classifier that, if $\mathcal{E}_\epsilon$ holds, uses LDA with $\hat{\mu}_j$, $\hat{\Sigma}_j$ and $\hat{\beta}$, and, if $\mathcal{E}_\epsilon$ fails, makes a random guess. Now, AUC for this classifier differs from $\widetilde{\mathrm{AUC}}$ by at most $p(\epsilon)$. Hence,

$$\begin{aligned}\left|\widetilde{\mathrm{AUC}} - \mathrm{AUC}\right| &\le \max\left\{\left|\inf_\epsilon \mathrm{AUC}^{\mathrm{pert}} - \mathrm{AUC}\right|,\right.\\ &\qquad \left.\left|\sup_\epsilon \mathrm{AUC}^{\mathrm{pert}} - \mathrm{AUC}\right|\right\} + p(\epsilon)\\ &\le C\epsilon + p(\epsilon)\ ,\end{aligned} \tag{17}$$

where C is as in (16). By (9), (13) and (14), for each fixed $\epsilon \in (0, \epsilon_0]$, $p(\epsilon) \to 0$ as the training sample sizes increase. Letting first those sample sizes diverge, and then $\epsilon \to 0$, in (17), we see that (8) and (12) hold as the training sample sizes increase.

Note that this argument does not actually need the training data to be Gaussian. Only the test data value $\mathbf{x}$ should have that property; it should satisfy $\mathbf{x} \sim \mathrm{N}(\mu_j, \Sigma_j)$ if it comes from $\mathcal{C}_j$, for $j = 1$ or 2.

ACKNOWLEDGMENTS

This work was partly supported by an International Travel Grant to J.-H. Xue from the Royal Society of London. Thanks to the University of Melbourne for hosting J.-H. Xue's sabbatical in 2012. Thanks to the reviewers for their constructive comments leading to Sections 4.4, 4.5 and 5.

REFERENCES

[1] N. V. Chawla, N. Japkowicz, and A. Kotcz, "Editorial: Special issue on learning from imbalanced data sets," *SIGKDD Explorations*, vol. 6, no. 1, pp. 1–6, 2004.

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

[3] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 9, pp. 1263–1284, Sep. 2009.

[4] G. M. Weiss and F. Provost, "Learning when training data are costly: The effect of class distribution on tree induction," *J. Artif. Intell. Res.*, vol. 19, pp. 315–354, 2003.

[5] J. G. Xie and Z. D. Qiu, "The effect of imbalanced data sets on LDA: A theoretical and empirical analysis," *Pattern Recognit.*, vol. 40, no. 2, pp. 557–562, 2007.

[6] J.-H. Xue and D. M. Titterington, "Do unbalanced data have a negative effect on LDA?" *Pattern Recognit.*, vol. 41, no. 5, pp. 1558–1571, 2008.

[7] J. Van Hulse and T. Khoshgoftaar, "Knowledge discovery from imbalanced and noisy data," *Data Knowl. Eng.*, vol. 68, no. 12, pp. 1513–1542, 2009.

[8] E. S. Gilbert, "The effect of unequal variance-covariance matrices on Fisher's linear discriminant function," *Biometrics*, vol. 25, no. 3, pp. 505–515, 1969.

[9] S. Marks and O. J. Dunn, "Discriminant function when covariance matrices are unequal," *J. Amer. Statist. Assoc.*, vol. 69, no. 345, pp. 555–559, 1974.

[10] J. A. Hanley and B. J. McNeil, "The meaning and use of the area under a receiver operating characteristic (ROC) curve," *Radiology*, vol. 143, no. 1, pp. 29–36, 1982.

[11] D. J. Hand and R. J. Till, "A simple generalisation of the area under the ROC curve for multiple class classification problems," *Mach. Learn.*, vol. 45, no. 2, pp. 171–186, 2001.

[12] F. E. Harrell, *Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis*. New York, NY, USA: Springer, 2001.

[13] M. S. Pepe, *The Statistical Evaluation of Medical Tests for Classification and Predictions*, Oxford, U.K.: Oxford Univ. Press, 2003.

[14] S. Ma and J. Huang, "Penalized feature selection and classification in bioinformatics," *Briefings Bioinformat.*, vol. 9, no. 5, pp. 392–403, 2008.

[15] J. Q. Su and J. S. Liu, "Linear combinations of multiple diagnostic markers," *J. Amer. Statist. Assoc.*, vol. 88, no. 424, pp. 1350–1355, 1993.

[16] A. Liu, E. F. Schisterman, and Y. Zhu, "On linear combinations of biomarkers to improve diagnostic accuracy," *Statist. Med.*, vol. 24, no. 1, pp. 37–47, 2005.

[17] L. Breiman, J. Friedman, R. Olshen, and C. Stone, *Classification and Regression Trees*. Belmont, CA, USA: Wadsworth, 1984.

[18] C. Elkan, "The foundations of cost-sensitive learning," in *Proc. 17th Int. Joint Conf. Artif. Intell.*, 2001, vol. 2, pp. 973–978.