A New Strategy of Cost-Free Learning in the Class Imbalance Problem

Xiaowan Zhang and Bao-Gang Hu, *Senior Member, IEEE*

Abstract—In this work, we define cost-free learning (CFL) formally in comparison with cost-sensitive learning (CSL). The main difference between them is that a CFL approach seeks optimal classification results without requiring any cost information, even in the class imbalance problem. In fact, several CFL approaches exist in the related studies, such as sampling and some criteria-based approaches. However, to our best knowledge, none of the existing CFL and CSL approaches are able to process the abstaining classifications properly when no information is given about errors and rejects. Based on information theory, we propose a novel CFL which seeks to maximize normalized mutual information of the targets and the decision outputs of classifiers. Using the strategy, we can handle binary/multi-class classifications with/without abstaining. Significant features are observed from the new strategy. While the degree of class imbalance is changing, the proposed strategy is able to balance the errors and rejects accordingly and automatically. Another advantage of the strategy is its ability of deriving optimal rejection thresholds for abstaining classifications and the "*equivalent*" costs in binary classifications. The connection between rejection thresholds and ROC curve is explored. Empirical investigation is made on several benchmark data sets in comparison with other existing approaches. The classification results demonstrate a promising perspective of the strategy in machine learning.

Index Terms—Classification, class imbalance, cost-free learning, cost-sensitive learning, abstaining, mutual information, ROC

1 Introduction

Imbalanced data sets [1], [2] arise frequently in a variety of real-world applications, such as medicine, biology, finance, and computer vision. Generally, users focus more on the *minority* class and consider the cost of misclassifying a minority class to be more expensive. Unfortunately, most *conventional classification* algorithms assume that the class distributions are balanced or the misclassification costs are equal. They seek to maximize the overall accuracy which yet cannot distinguish the error types. Therefore, they may neglect the significance of the minority class and tend toward the majority class. Learning in the class imbalance is thus of high importance in data mining and machine learning.

From the background of this problem, various methods are developed within a category called *cost-sensitive learning* (CSL), such as costs to test [3], to relabel training instances [4], to sample [5], to weight instances [6], and to find a decision threshold [7], [8]. These methods use unequal costs to make a bias toward the minority class. Generally, when the costs are not given, these methods cannot work properly. A comprehensive review of learning in the class imbalance problem is provided by He and Garcia [9].

When there exist some uncertainties in the decision, it may be better to apply *abstaining classification* [10] to reduce the chance of a potential misclassification. Significant benefits have been obtained from abstaining classification, particularly in very critical applications [11], [12]. The optimal rejection thresholds could be found through minimizing a loss function in a cost-sensitive setting [13], [14], [15]. The possibility of designing loss functions for classifiers with a reject option is also explored [16]. In the context of abstaining classifications, the existing CSL approaches require the cost terms associated to the rejects. However, one often fails to provide such information (e.g., the costs of not making decisions in disease diagnosis). Up to now, there seems no proper guideline to give the information in terms of the skew ratio. Obviously, a reject option adds another degree of complexity in classifications over the non-abstaining approaches. For advancing the technology and being compatible with human intelligence, we consider the abstaining strategy will become a common option for most learning machines in future.

In the class imbalance problem, CSL is an important research direction. Based on the definition in [17], we extend it below by including the situation of abstaining.

Definition 1. Cost-Sensitive Learning *is a type of learning that takes the misclassification costs and/or rejection costs into consideration. The goal of this type of learning is to minimize the total cost.*

CSL generally requires modelers or users to specify cost terms for reaching the goal. However, this work addresses one open issue which is mostly overlooked:

"How to conduct a learning in the class imbalance problem when costs are unknown for errors and rejects"?

In fact, the issue is not unusual in real-world applications (e.g., Maloof [18]; Zadrozny and Elkan [19]). Therefore, we propose another category of learning below for distinguishing the differences between the present work and the existing studies in CSL.

- *The authors are with the National Laboratory of Pattern Recognition, Institute of Automation Chinese Academy of Sciences, Beijing 100190, P.R. China. E-mail: {xwzhang, hubg}@nlpr.ia.ac.cn.*

Manuscript received 21 July 2013; revised 15 Feb. 2014; accepted 12 Mar. 2014. Date of publication 17 Mar. 2014; date of current version 29 Oct. 2014.
Recommended for acceptance by L. Khan.
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/TKDE.2014.2312336

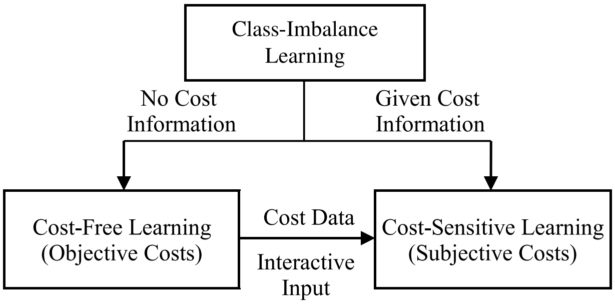


Fig. 1. Cost-free learning and cost-sensitive learning.

Definition 2. Cost-Free Learning (CFL) *is a type of learning that does not require the cost terms associated with the misclassifications and/or rejects as the inputs. The goal of this type of learning is to get optimal classification results without using any cost information.*

Any classification that does not involve costs can be considered as CFL. It may have its own optimization strategy, such as maximizing AUC or the overall accuracy. In this case, those conventional classifications that aim to maximize the overall accuracy can be categorized as CFL, although they are not doing well in the class imbalance problem. It is understandable that CFL may face a bigger challenge which is shown by the fact that most existing approaches may fail to present reasonable solutions to the open issue. This work attempts to provide an applicable learning strategy in CFL.

We extend Hu's [20] study on mutual information classifiers. While Hu presents the theoretical formulas, no learning approaches and results are shown for the real-world data sets. Hence, this work focuses on learning and presents main contributions as follows:

- We propose a CFL strategy in the class imbalance problem. Using *normalized mutual information* (NI) as the learning target, we conduct the learning from cost-insensitive classifiers. Therefore, we are able to adopt conventional classifiers for simple and direct implementations. The most advantage of this strategy is its unique feature in classification scenarios where one has no knowledge of costs.
- We study the relations between the strategy and some existing approaches. First, we derive the "*equivalent*" costs and the rejection thresholds for binary classifications by using the strategy. The costs are "*objective*", for they are purely determined by the distributions of the given data sets. They can be taken as useful references for "*subjective*" cost specifications in CSL (Fig. 1). Second, we present graphical interpretations of ROC curve plots for both non-abstaining and abstaining classifiers. From the plots, the intrinsic differences between the strategy and other existing approaches are explained in the cases when one class becomes extremely rare.
- We conduct empirical studies on binary class and multi-class problems. Specific investigation is made on abstaining classifications, and we obtain several results on benchmark data sets. The results confirm the advantages of the strategy and show the promising perspective of CFL in imbalanced data sets.

1.1 Related Work

When costs are unequal and unknown, Maloof [18] uses ROC curve to show the performance of binary classifications under different cost settings. To make fair comparisons, an alternative to AUC is proposed to evaluate different classifiers under the same cost ratio distribution [21]. These studies can be viewed as comparing classifiers rather than finding optimal operating points. Cost curve [14], [22] can be used to visualize optimal expected costs over a range of cost settings, but it does not suit multi-class problem. Zadrozny and Elkan [19] apply least-squares multiple linear regression to estimate the costs. The method requires cost information of the training sets to be known. Cross validation (CV) [23] is proposed to choose from a limited set of cost values, and the final decisions are made by users.

There exists some CFL approaches in the class imbalance problem. Various sampling strategies [24], [25], [26] try to modify the imbalanced class distributions. Active learning [27] is also investigated to select desired instances and the feature selection techniques [28], [29] are applied to combat the class imbalance problem for high-dimensional data sets. Besides, ensemble learning methods [30], [31] are used to improve the generalization of predicting the minority class. To reduce the influence of imbalance, Hellinger distance is applied to decision trees as a splitting criterion for its skew insensitivity [32]. And the recognition-based methods [33], [34] that train on a single class are proposed as alternatives to the discrimination-based methods. Sun et al. [35] get costs through maximizing *geometric mean* (G-mean) or *F-measure* which has the ability of balancing the performance of each class. However, all CFL methods above do not take abstaining into consideration and may fail to process the abstaining classifications.

In regards to abstaining classification, some strategies have been proposed for defining optimal reject rules. Pietraszek [36] proposes a bounded-abstaintion model with ROC analysis, and Fumera et al. [37] seek to maximize accuracy while keeping the reject rate below a given value. However, the bound information and the targeted reject rate are required to be specified respectively. When there is no prior knowledge of these settings, it is hard to determine the values. Li and Sethi [38] restrict the maximum error rate of each class, but the rates may conflict when they are arbitrarily given.

1.2 Paper Organization

The remainder of this paper is organized as follows: In Section 2, a brief review of NI is provided. We present our CFL strategy in Section 3. Section 4 analyzes the relations between the optimal parameters and the cost terms, and presents the graphical interpretations of ROC curve plots. The experimental results are presented in Section 5. Finally, we conclude this work in Section 6.

2 Review: Normalized Mutual Information

Normalized mutual information has been used as an evaluation criterion to measure the degree of dependence between the targets T and the decision outputs Y, and it is denoted as

TABLE 1
Confusion Matrix C in m-Class Abstaining Classification

	Y				
T	1	2	$\dots$	m	$m+1$
1	c_{11}	c_{12}	$\dots$	c_{1m}	$c_{1(m+1)}$
2	c_{21}	c_{22}	$\dots$	c_{2m}	$c_{2(m+1)}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
m	c_{m1}	c_{m2}	$\dots$	c_{mm}	$c_{m(m+1)}$

$$NI(T,Y) = \frac{I(T,Y)}{H(T)},$$

where $I(T,Y)$ is the mutual information of two random variables T and Y, $H(T)$ is the Shannon's entropy of T. Note that $NI(T,Y)$ is in the range $[0,1]$.

Suppose an m-class abstaining classification, with each class denoted as $1,2,\dots,m$, and the rejected class denoted as $m+1$. The value of the target variable T ranges from 1 to m, while the decision output variable Y ranges from 1 to $m+1$. Then we have

$$I(T,Y) = \sum_{i=1}^{m}\sum_{j=1}^{m+1} P(T=i,Y=j)\log_2 \frac{P(T=i,Y=j)}{P(T=i)P(Y=j)},$$

$$H(T) = -\sum_{i=1}^{m} P(T=i)\log_2 P(T=i).$$

In general, as the exact probability distribution functions of T and Y are hard to derive, Hu et al. [39] apply empirical estimations to compute NI based on the confusion matrix. Table 1 illustrates an augmented confusion matrix C in an m-class abstaining classification by adding the last column as a rejected class $m+1$. The rows correspond to the states of the targets T, and the columns correspond to the states of the decision outputs Y. c_{ij} represents the number of the instances that belong to the ith class classified as the jth class, $i=1,2,\dots,m$, $j=1,2,\dots,m+1$. Nevertheless, the value of NI may be unchanged when rejects occur only in one class. In this situation, NI cannot distinguish rejects. It is proved that the modified form of NI that computes Y from 1 to m cannot only overcome this weakness but also has the good features of the original NI [39]. Therefore, the modified form is applied as the formula of NI for both non-abstaining and abstaining classifications:

$$\begin{aligned} NI(T,Y) &= \frac{\sum_{i=1}^{m}\sum_{j=1}^{m} P_e(T=i,Y=j)\log_2 \frac{P_e(T=i,Y=j)}{P_e(T=i)P_e(Y=j)}}{-\sum_{i=1}^{m} P_e(T=i)\log_2 P_e(T=i)} \\ &= -\frac{\sum_{i=1}^{m}\sum_{j=1}^{m} c_{ij}\log_2\left(\frac{c_{ij}}{C_i\sum_{i=1}^{m}(\frac{c_{ij}}{n})}\right)}{\sum_{i=1}^{m} C_i\log_2(\frac{C_i}{n})}, \end{aligned} \tag{1}$$

where Y is counted from 1 to m rather than to $m+1$. The subscript "e" is given for denoting empirical terms, $C_i = \sum_{j=1}^{m+1} c_{ij}$ is the total number of instances in the ith class, $i=1,2,\dots,m$, and $n=\sum_{i=1}^{m}\sum_{j=1}^{m+1} c_{ij}$ is the total number in the confusion matrix. Besides overcoming the limitation of the original NI, (1) can simplify the computation in the abstaining classification and unify the forms of computations in both non-abstaining and abstaining classifications.

Principe et al. [40] present a schematic diagram of *information theory learning* (ITL) and they mention that maximizing mutual information as the target function makes the decision outputs correlate with the targets as much as possible. Recently, a study [20] confirms that ITL opens a new perspective for classifier design. MacKay [41] recommends mutual information for its single rankable value which makes more sense than error rate. Hu et al. [39] study theoretically for the first time on both error types and reject types in binary classifications. They consider information-theoretic measures most promising in providing "*objectivity*" to classification evaluations in class imbalance problems. The above viewpoints of mutual information motivate our following NI-based strategy for CFL in the class imbalance problem.

3 NI-Based Classification

In this work, we distinguish two types of classifications, namely, "*non-abstaining classification*" for no rejection and "*abstaining classification*" for rejection. From the phenomenon that different error types and reject types produce different effects on NI, one can derive a conclusion that NI considers the costs to be unequal, unlike accuracy. In fact, the cost information is hiding in NI, and we take advantage of its bias toward the minority class. The bias can be changed through moving the decision thresholds, and the value of NI is changed accordingly. We focus our study on the probabilistic classifiers in the present work, although it can also be applied to non-probabilistic classifiers [42].

Let $\boldsymbol{x} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_n]^T$ denote a data matrix with n instances to be classified, $\boldsymbol{x}_l \in \mathbb{R}^d$ is the input feature vector, $l=1,2,\dots,n$. The target vector is denoted as $\boldsymbol{t} = [t_1, t_2, \dots, t_n]^T$, $t_l \in T = \{1,2,\dots,m\}$. The decision output vector is denoted as $\boldsymbol{y} = [y_1, y_2, \dots, y_n]^T$, $y_l \in Y = \{1,2,\dots,m\}$ for non-abstaining classification while $y_l \in Y = \{1,2,\dots,m+1\}$ for abstaining classification. Then for both non-abstaining and abstaining classifications, we have a generalized formula with NI being a function of the data set and the decision thresholds:

$$\begin{aligned} &NI = NI\big(\boldsymbol{t}, \boldsymbol{y} = f(\boldsymbol{\varphi}(\boldsymbol{x}), \boldsymbol{\tau})\big), \\ &y_l = \begin{cases} \arg\max\limits_i\left(\dfrac{\varphi_i(\boldsymbol{x}_l)}{\tau_i}\right) & \text{if } \max\left(\frac{\varphi_i(\boldsymbol{x}_l)}{\tau_i}\right) \ge 1, \\ m+1 & \text{otherwise}, \end{cases} \\ &0 < \tau_i \le 1,\ i=1,2,\dots,m,\ l=1,2,\dots,n, \end{aligned} \tag{2}$$

where $\boldsymbol{\varphi}(\boldsymbol{x}) \in \mathbb{R}^{n\times m}$ denotes the real-value output matrix of a probabilistic classifier for n instances, $\varphi_i(\boldsymbol{x}_l)$ is the probabilistic output of class i for $\boldsymbol{x}_l$, $\sum_{i=1}^{m}\varphi_i(\mathrm{x}_l) = 1$ and $0 \le \varphi_i(\boldsymbol{x}_l) \le 1$. $\boldsymbol{\tau} = [\tau_1, \tau_2, \dots, \tau_m]^T \in \mathbb{R}^m$ is the vector parameter of the decision thresholds. The decision rule of y_l is proposed in this form to avoid classifying an instance $\boldsymbol{x}_l$ into more than one class.

3.1 Non-Abstaining Classification

In non-abstaining classification, the first condition for deriving y_l in (2) should only be satisfied, i.e.,

$$\begin{aligned} &y_l = \arg\max_i \left(\frac{\varphi_i(\boldsymbol{x}_l)}{\tau_i}\right), 0 < \tau_i \leq 1, \\ &i = 1, 2, \ldots, m, l = 1, 2, \ldots, n. \end{aligned}$$

Here we introduce α_i as the weight parameter for $\varphi_i(\boldsymbol{x}_l)$, since we need it in Section 4.2. Let $\phi_i(\boldsymbol{x}_l) = \alpha_i \varphi_i(\boldsymbol{x}_l)$, $\alpha_i = \frac{\tau_m}{\tau_i}$ and $\alpha_m = 1$, then we have the following:

$$\begin{aligned} \phi_i(\boldsymbol{x}_l) &= \alpha_i \varphi_i(\boldsymbol{x}_l) \\ &= \tau_m \frac{\varphi_i(\boldsymbol{x}_l)}{\tau_i}. \end{aligned}$$

It is obvious that $\arg\max_i \phi_i(\boldsymbol{x}_l) = \arg\max_i \left(\frac{\varphi_i(\boldsymbol{x}_l)}{\tau_i}\right)$, and the optimal decision for y_l remains the same. The effect of assigning weights to the probabilistic outputs is the same as setting decision thresholds. Therefore, we denote $\boldsymbol{\alpha} = [\alpha_1, \ \alpha_2, \ldots, 1]^T \in \mathbb{R}^m$ as the weight parameter vector, and the class assignment rule for $y_l = f(\boldsymbol{\varphi}(\boldsymbol{x}_l), \boldsymbol{\alpha})$ is based on the highest weighted probabilistic outputs. For non-abstaining classification, we propose

$$\begin{aligned} &\text{maximize } NI\big(\boldsymbol{t}, \boldsymbol{y} = f(\boldsymbol{\varphi}(\boldsymbol{x}), \boldsymbol{\alpha})\big), \\ &\text{subject to} \\ &\quad y_l = \arg\max_i \alpha_i \varphi_i(\boldsymbol{x}_l), \\ &\quad \alpha_i > 0, \ i = 1, 2, \ldots, m, \ l = 1, 2, \ldots, n. \end{aligned} \tag{3}$$

In order to maximize NI, the optimal weight parameter $\boldsymbol{\alpha}^*$ should be

$$\boldsymbol{\alpha}^* = \arg \max_{\boldsymbol{\alpha}} NI\big(\boldsymbol{t}, \boldsymbol{y} = f(\boldsymbol{\varphi}(\boldsymbol{x}), \boldsymbol{\alpha})\big). \tag{4}$$

3.2 Abstaining Classification

We denote $\boldsymbol{T}_r = [T_{r1}, T_{r2}, \ldots, T_{rm}]^T \in \mathbb{R}^m$ as the rejection threshold vector in dealing with abstaining classificaiton. Let $1 - T_{ri} = \tau_i$, T_{ri} is in the range $[0, 1)$, $i = 1, 2, \ldots, m$. The decision output for $y_l = f(\boldsymbol{\varphi}(\boldsymbol{x}_l), \boldsymbol{T}_r)$ lies within $m + 1$ classes. Then we propose

$$\begin{aligned} &\text{maximize } NI\big(\boldsymbol{t}, \boldsymbol{y} = f(\boldsymbol{\varphi}(\boldsymbol{x}), \boldsymbol{T}_r)\big), \\ &\text{subject to} \\ &y_l = \begin{cases} \arg\max_i \left(\dfrac{\varphi_i(\boldsymbol{x}_l)}{1 - T_{ri}}\right) & \text{if } \max\left(\frac{\varphi_i(\boldsymbol{x}_l)}{1 - T_{ri}}\right) \geq 1, \\ m + 1 & \text{otherwise}, \end{cases} \\ &0 \leq T_{ri} < 1, 0 \leq \sum_{i=1}^{m} T_{ri} < m - 1, \\ &i = 1, 2, \ldots, m, \ l = 1, 2, \ldots, n. \end{aligned} \tag{5}$$

Note that $m - 1$ is the loose upper bound for the summation $\sum_{i=1}^m T_{ri}$. Assume a situation that all instances satisfy the first condition in (5), and $\frac{\varphi_i(\boldsymbol{x}_l)}{1 - T_{ri}} \geq 1$ for all probabilistic outputs, i.e., $\forall i, l, \ \varphi_i(\boldsymbol{x}_l) \geq 1 - T_{ri}$. Then we get the following:

$$\begin{aligned} \sum_{i=1}^{m} \varphi_i(\boldsymbol{x}_l) &\geq \sum_{i=1}^{m} (1 - T_{ri}), \\ \sum_{i=1}^{m} T_{ri} &\geq m - 1. \end{aligned}$$

If $\sum_{i=1}^m T_{ri}$ falls in this interval, the condition of rejection would never be satisfied and the proposal of abstaining classification is ineffective. Reversely, this extreme situation would not happen if $\sum_{i=1}^m T_{ri} < m - 1$.

In order to maximize NI, the optimal rejection threshold vector ${\boldsymbol{T}_r}^*$ should be

$${\boldsymbol{T}_r}^* = \arg \max_{\boldsymbol{T}_r} NI\big(\boldsymbol{t}, \boldsymbol{y} = f(\boldsymbol{\varphi}(\boldsymbol{x}), \boldsymbol{T}_r)\big). \tag{6}$$

3.3 Optimization Algorithm

The present framework is proposed based on the confusion matrix from which we compute NI, but it is not differentiable and non-convex. We apply a general optimization algorithm called "*Powell Algorithm*" which is a direct method for nonlinear optimization without calculating the derivatives [43]. It is also widely used in image registration to find optimal registration parameters.

The algorithm is given in Algorithm 1 that finds $\boldsymbol{\tau}^*$ for demonstration purposes. We can also apply it to find both $\boldsymbol{\alpha}^*$ and $\mathbf{T}_\mathrm{r}^*$. For Step 6 and Step 10, we use *bracketing method* to find three starting points and use *Brent's Method* to realize one-dimensional optimization. W iterations of the basic procedure lead to $W(\mathcal{D} + 1)$ one-dimensional optimizations. One disadvantage of this algorithm is that it may find a local extrema. Hence, we randomly choose the starting points several times and then pick the best one. In non-abstaining binary classification, $\mathcal{D} = 1$, so we just work from Steps 4 to 7 once and assign the value of $\boldsymbol{\tau}_W^{(2)}$ to $\boldsymbol{\tau}^*$.

Algorithm 1 Learning algorithm

```
Input: Probabilistic outputs $\boldsymbol{\varphi}(\boldsymbol{x})$, target labels $\boldsymbol{t}$, $\mathcal{D}$ as the degree of
    freedom in $\boldsymbol{\tau}$.
Output: $\boldsymbol{\tau}^*$
 1: Initialize $\boldsymbol{\tau_1}$ as a random vector in the range of $\boldsymbol{\tau}$, $\boldsymbol{d_1}, \boldsymbol{d_2}, \ldots, \boldsymbol{d_\mathcal{D}}$
    as linear independent vectors, number of iterations $W = 0$, $\varepsilon \geq 0$.
 2: Iterative Search Phase:
 3: repeat
 4:   $W = W + 1$. Let $\boldsymbol{\tau_W^{(1)}} = \boldsymbol{\tau_W}$.
 5:   for each direction $\boldsymbol{d_i}$, $i = 1$ to $\mathcal{D}$ do
 6:     $\bar{\eta}^{(i)} = \arg\max_{\eta \in \mathbb{R}} NI\big(\boldsymbol{t}, \boldsymbol{y} = f(\boldsymbol{\varphi}(\boldsymbol{x}), \boldsymbol{\tau_W^{(i)}} + \eta \boldsymbol{d_i})\big)$;
 7:     Update $\boldsymbol{\tau_W}$ in the current direction: $\boldsymbol{\tau_W^{(i+1)}} = \boldsymbol{\tau_W^{(i)}} + \bar{\eta}^{(i)} \boldsymbol{d_i}$;
 8:   end for
 9:   Update the directions: $\boldsymbol{d_i} = \boldsymbol{d_{i+1}}, i = 1, 2, \ldots, \mathcal{D} - 1$;
                             $\boldsymbol{d_\mathcal{D}} = \boldsymbol{\tau_W^{(\mathcal{D}+1)}} - \boldsymbol{\tau_W}$;
10:   $\eta_W^* = \arg\max_{\eta \in \mathbb{R}} NI\big(\boldsymbol{t}, \boldsymbol{y} = f(\boldsymbol{\varphi}(\boldsymbol{x}), \boldsymbol{\tau_W} + \eta \boldsymbol{d_\mathcal{D}})\big)$;
11:   Update $\boldsymbol{\tau}$ after the current iteration: $\boldsymbol{\tau_{W+1}} = \boldsymbol{\tau_W} + \eta_W^* \boldsymbol{d_\mathcal{D}}$;
12: until $||\boldsymbol{\tau_{W+1}} - \boldsymbol{\tau_W}||_2 \leq \varepsilon$
13: Return $\boldsymbol{\tau}^* = \boldsymbol{\tau_{W+1}}$.
```

4 Relations in Binary Classification

The previous section completes the essence of the present framework. It can be regarded as a generic way to make the conventional learning algorithms information-based.

The optimal parameters reflect the degree of bias implied by NI, and may reveal the cost information to some extent. In this section, we focus on binary classification and analyze the relations between the optimal parameters and the cost terms. Moreover, we discover some graphical interpretations of performance measures on ROC curve, which allows the users to adjust the parameters more conveniently using ROC curve.

4.1 Normalized Cost Matrix

Friedel et al. [14] derive normalized cost matrix based on the *overall risk* which is written as

$$Risk = \sum_{i,j} \lambda_{ij} p(j \,|\, i) p(i), \tag{7}$$

where λ_{ij} is the original cost in the common cost matrix that assigns an instance of class i to class j, $p(j \,|\, i)$ is the true probability in such situation, and $p(i)$ is the true prior probability of class i. The *conditional risk* of assigning an instance $\boldsymbol{x}_l$ to class j is

$$Risk(j \,|\, \boldsymbol{x}_l) = \sum_{i=1}^{m} \lambda_{ij} p(i \,|\, \boldsymbol{x}_l), \tag{8}$$

where $p(i \,|\, \boldsymbol{x}_l)$ is the true posterior probability of class i given $\boldsymbol{x}_l$. By applying the way of transforming costs [14], we find that the normalization way for the overall risk is also applicable for the conditional risk.

In binary classification, we refer to classes 1 and 2 as *negative class* (N) and *positive class* (P), respectively. We denote λ_{FN}, λ_{FP}, λ_{TN}, λ_{TP}, λ_{RN} and λ_{RP} to be the costs of *false negative*, *false positive*, *true negative*, *true positive*, *reject negative*, and *reject positive*, respectively. Therefore, the normalized cost matrix for non-abstaining binary classification can be denoted as

$$\overline{\lambda}_{no\_rej} = \begin{bmatrix} \overline{\lambda}_{TN} & \overline{\lambda}_{FP} \\ \overline{\lambda}_{FN} & \overline{\lambda}_{TP} \end{bmatrix} = \begin{bmatrix} 0 & \overline{\lambda}_{FP} \\ 1 & 0 \end{bmatrix} \tag{9}$$

with $\beta = \lambda_{FN} - \lambda_{TP}$, then $\overline{\lambda}_{TN} = \frac{\lambda_{TN}-\lambda_{TN}}{\beta} = 0$, $\overline{\lambda}_{FP} = \frac{\lambda_{FP}-\lambda_{TN}}{\beta}$, $\overline{\lambda}_{FN} = \frac{\lambda_{FN}-\lambda_{TP}}{\beta} = 1$, $\overline{\lambda}_{TP} = \frac{\lambda_{TP}-\lambda_{TP}}{\beta} = 0$.

Similarly, the normalized cost matrix for abstaining binary classification can be denoted as

$$\overline{\lambda}_{rej} = \begin{bmatrix} \overline{\lambda}_{TN} & \overline{\lambda}_{FP} & \overline{\lambda}_{RN} \\ \overline{\lambda}_{FN} & \overline{\lambda}_{TP} & \overline{\lambda}_{RP} \end{bmatrix} = \begin{bmatrix} 0 & \overline{\lambda}_{FP} & \overline{\lambda}_{RN} \\ 1 & 0 & \overline{\lambda}_{RP} \end{bmatrix} \tag{10}$$

with $\overline{\lambda}_{TN} = 0$, $\overline{\lambda}_{FP} = \frac{\lambda_{FP}-\lambda_{TN}}{\beta}$, $\overline{\lambda}_{RN} = \frac{\lambda_{RN}-\lambda_{TN}}{\beta}$, $\overline{\lambda}_{FN} = 1$, $\overline{\lambda}_{TP} = 0$, $\overline{\lambda}_{RP} = \frac{\lambda_{RP}-\lambda_{TP}}{\beta}$, $\beta = \lambda_{FN} - \lambda_{TP}$. The first two columns contain the misclassification costs, while the last column indicates the rejection costs.

It is reasonable to assume that the values of the original correct classification costs and misclassification costs in the common cost matrix are not affected by introducing a reject option. Therefore, what is noteworthy is that $\overline{\lambda}_{FP}$ in (9) is consistent with that in (10).

4.2 Optimal Weight and Misclassification Cost

In non-abstaining binary classification, it is feasible to set the decision thresholds as $\boldsymbol{\tau} = [1 - \tau_P, \tau_P]^T$, which has one degree of freedom.

Under the class assignment rule of minimizing the conditional risk in (8), the relation between the decision thresholds and the costs has been derived by Elkan [7]. Considering the normalized cost matrix in (9), then the decision threshold τ_P^* of the positive class for making optimal decision is

$$\tau_P^* = \frac{\overline{\lambda}_{FP}}{1 + \overline{\lambda}_{FP}}, \tag{11}$$

with the misclassification cost $\overline{\lambda}_{FP}$ be the variable. However, it is hard to decide the value of $\overline{\lambda}_{FP}$.

In our present work, the optimal weight vector is $\boldsymbol{\alpha}^* = [\alpha_N^*, 1]^T$. According to the class assignment rule based on the highest weighted posterior probability, the optimal prediction is the positive class if and only if $\alpha_N^* p(N \,|\, \boldsymbol{x}_l) \leq p(P \,|\, \boldsymbol{x}_l)$, i.e., $\alpha_N^*(1 - p(P \,|\, \boldsymbol{x}_l)) \leq p(P \,|\, \boldsymbol{x}_l)$. Hence, the decision threshold τ_P^{**} of the positive class for making optimal decision is

$$\tau_P^{**} = \frac{\alpha_N^*}{1 + \alpha_N^*}. \tag{12}$$

Suppose that the two class assignment rules above share the same decision thresholds, then (11) and (12) should be equal, i.e., $\overline{\lambda}_{FP} = \alpha_N^*$. The value of the misclassification cost can thus be decided, and we give the following definition:

Definition 3. *Given the optimal weight α_N^*, the* "equivalent" misclassification cost *is defined as*

$$\overline{\lambda}_{FP} = \alpha_N^*. \tag{13}$$

The minority class is usually regarded as the positive class, and it is assumed that $\overline{\lambda}_{FP} < \overline{\lambda}_{FN}$, i.e., $\overline{\lambda}_{FP} < 1$. This facilitates the users to verify whether the *"equivalent"* misclassification cost agrees with human assumption by comparing α_N^* with 1.

4.3 Optimal Rejection Thresholds and Costs

In abstaining binary classification, the relations between the rejection thresholds and the costs can be presented in a form of explicit formulae [20]. With the optimal rejection threshold vector $\mathbf{T_r^*} = [T_{rN}^*, T_{rP}^*]^T$ and the normalized cost matrix in (10), these relations are

$$T_{rN}^* = \frac{\overline{\lambda}_{RN}}{1 + \overline{\lambda}_{RN} - \overline{\lambda}_{RP}},$$
$$T_{rP}^* = \frac{\overline{\lambda}_{RP}}{\overline{\lambda}_{FP} - \overline{\lambda}_{RN} + \overline{\lambda}_{RP}},$$

which imply a parameter redundancy. In addition, the value of $\overline{\lambda}_{FP}$ derived from (13) can be utilized as a prior knowledge under the assumption of cost consistency.

Definition 4. *Given the* "equivalent" *misclassification cost $\overline{\lambda}_{FP} = \alpha_N^*$, the* "equivalent" rejection costs *are defined as*

$$\begin{aligned} \overline{\lambda}_{RN} &= \frac{T_{rN}^*(1 - T_{rP}^*) - T_{rN}^* T_{rP}^* \overline{\lambda}_{FP}}{1 - T_{rN}^* - T_{rP}^*}, \\ \overline{\lambda}_{RP} &= \frac{-T_{rN}^* T_{rP}^* + (1 - T_{rN}^*) T_{rP}^* \overline{\lambda}_{FP}}{1 - T_{rN}^* - T_{rP}^*}. \end{aligned} \tag{14}$$

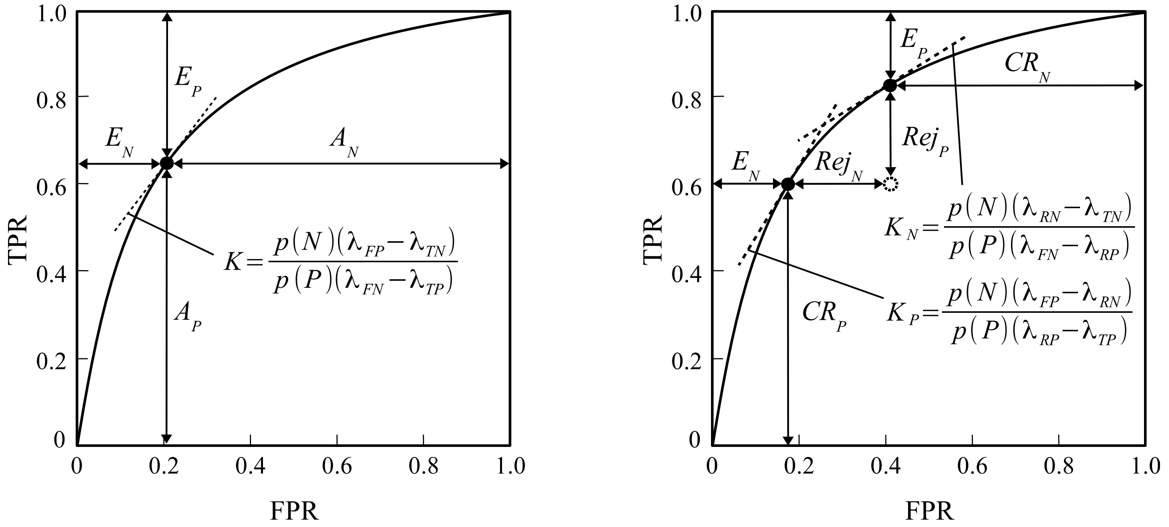
Based on [20], one can have the relations $\overline{\lambda}_{TN} < \overline{\lambda}_{RN} < \overline{\lambda}_{FP}$ and $\overline{\lambda}_{TP} < \overline{\lambda}_{RP} < \overline{\lambda}_{FN}$. Then we can obtain the following properties from (14):

P1. If $0 < \overline{\lambda}_{RN} < \overline{\lambda}_{FP}$, we have $0 < T_{rN}^* < \frac{\alpha_N^*}{1+\alpha_N^*}$ and $T_{rP}^* < \frac{1}{1+\alpha_N^*}$;

P2. If $0 < \overline{\lambda}_{RP} < 1$, we have $T_{rN}^* < \frac{\alpha_N^*}{1+\alpha_N^*}$ and $0 < T_{rP}^* < \frac{1}{1+\alpha_N^*}$;

P3. If $0 < T_{rN}^* < \frac{\alpha_N^*}{1+\alpha_N^*}$ and $0 < T_{rP}^* < \frac{1}{1+\alpha_N^*}$, then $0 < \overline{\lambda}_{RN} < \overline{\lambda}_{FP}$ and $0 < \overline{\lambda}_{RP} < 1$.

The above properties facilitate the users to verify whether the *"equivalent"* rejection costs agree with the assumed relations from the optimal values of the weight and the rejection thresholds.



(a) For non-abstaining classification

(b) For abstaining classification

Fig. 2. Graphical interpretations of ROC curves. (a) For non-abstaining classification. (b) For abstaining classification.

4.4 Graphical Interpretations of ROC Curve Plots with/without Abstaining

In binary classification, an ROC curve plot presents complete information about the performance of each class [44], so that an *overall performance measure* [45], such as AUC, can be formed. This is a preferred feature in processing class imbalance problems [46]. Furthermore, an ROC curve can also provide the graphical interpretations for non-abstaining and abstaining classifications in Fig. 2, where TPR and FPR are *true positive rate* and *false positive rate*. We denote A, CR, E and Rej to be *accuracy*, *correct recognition rate*, *error rate*, and *reject rate*, respectively. C_N and C_P are the total numbers of the negatives and positives, respectively. $c_{FN}, c_{FP}, c_{TN}, c_{TP}, c_{RN}$ and c_{RP} are the numbers of the *false negatives*, *false positives*, *true negatives*, *true positives*, *reject negatives*, and *reject positives*, respectively. Their relations are shown as follows:

Non-abstaining:

$$\begin{aligned}&A_N+E_N=1, \text{and } A_P+E_P=1,\\&A_N=\frac{c_{TN}}{C_N}, E_N=\frac{c_{FP}}{C_N}, A_P=\frac{c_{TP}}{C_P}, E_P=\frac{c_{FN}}{C_P}.\end{aligned} \tag{15a}$$

Abstaining:

$$\begin{aligned}&CR_N+E_N+Rej_N=1, \text{and } CR_P+E_P+Rej_P=1,\\&CR_N=\frac{c_{TN}}{C_N}, E_N=\frac{c_{FP}}{C_N}, Rej_N=\frac{c_{RN}}{C_N},\\&CR_P=\frac{c_{TP}}{C_P}, E_P=\frac{c_{FN}}{C_P}, Rej_P=\frac{c_{RP}}{C_P}.\end{aligned} \tag{15b}$$

Several observations are summarized below for understanding the features of ROC plots. To begin with, we discuss an ROC curve in a non-abstaining classification, as shown in Fig. 2a. For a theoretical ROC curve which is concave, the classification decision is made by K, the *slope of ROC curve*, in the form of [47]:

$$K=\frac{p(N)}{p(P)}\frac{\lambda_{FP}-\lambda_{TN}}{\lambda_{FN}-\lambda_{TP}}=\frac{p(N)}{p(P)}\overline{\lambda}_{FP}, \tag{16}$$

which is also equivalent to the likelihood ratio [48]:

$$L=\frac{p(x\,|\,P)}{p(x\,|\,N)}=\frac{p(N)}{p(P)}\frac{\lambda_{FP}-\lambda_{TN}}{\lambda_{FN}-\lambda_{TP}}=\frac{p(N)}{p(P)}\overline{\lambda}_{FP}. \tag{17}$$

From (16), one can observe that

$$if\ p(P)\rightarrow 0, then\ K\rightarrow\infty, \tag{18a}$$

$$and\ E_P=1, A_P=0, E_N=0, A_N=1, \tag{18b}$$

TABLE 2
Description of the Data Sets

Data Set	#Inst	#Attr	#C	Class Distribution
Ism	11,180	7	2	10,920/260(=42.00)
Nursery(very_recom)	12,960	9	2	12,632/328(=38.51)
Letter(A)	20,000	17	2	19,211/789(=24.35)
Rooftop	17,829	10	2	17,048/781(=21.83)
Vehicle(opel)	846	19	2	634/212(=2.99)
Yeast(NUC)	1,484	10	2	1,055/429(=2.46)
Phoneme	5,404	6	2	3,818/1,586(=2.41)
German Credit	1,000	25	2	700/300(=2.33)
Diabetes	768	9	2	500/268(=1.87)
Gamma	19,020	11	2	12,332/6,688(=1.84)
Cardiotocography	2,126	22	3	1,655/295/176
Thyroid	7,200	22	3	6,666/368/166
Car	1,728	7	4	1,210/384/65/69
Pageblock	5,473	11	5	4,913/329/28/88/115

(#Inst: number of instances, #Attr: number of attributes, #C: number of classes).

TABLE 3
The Procedure of Our NI-Based Experiments

1. Apply stratified 3-fold cross validation on a data set. $\frac{2}{3}$ data belong to the training set and the remainder belong to the test set.
   a. Apply stratified 3-fold cross validation on the training set. $\frac{2}{3}$ data belong to the estimation set and the remainder belong to the validation set.
      i. Apply Algorithm 1 10 times to get the best parameter in each fold.
   b. Apply the mean value of 3 best parameters in step a to the training set.
   c. Predict the test set with the parameter obtained from step b.
2. Obtain the results of 3 test sets.

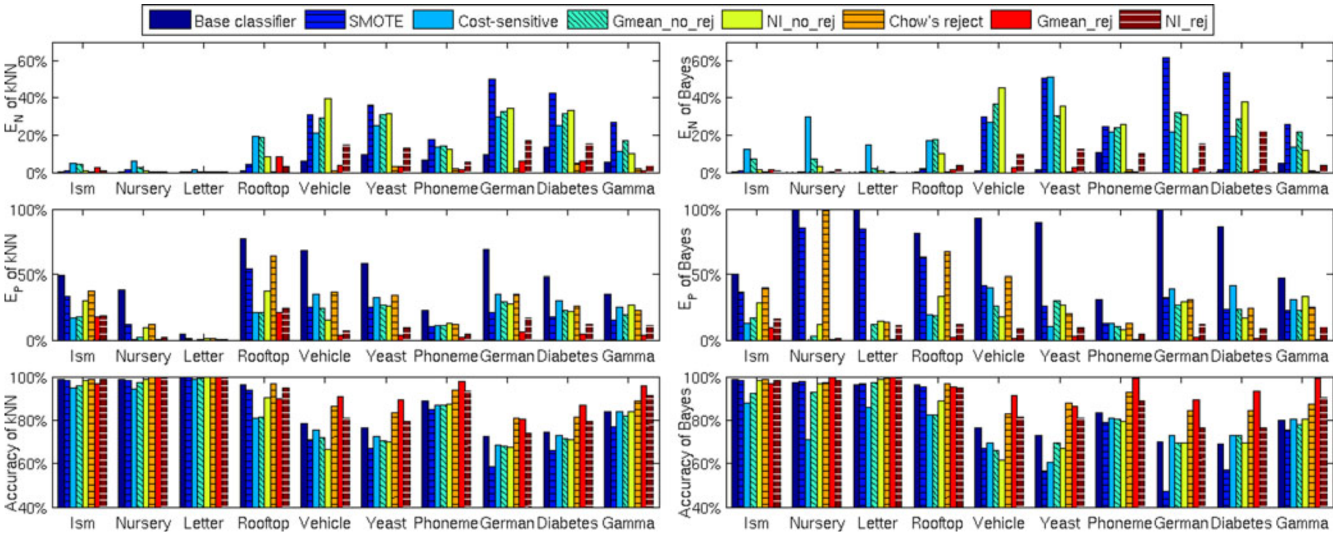


Fig. 3. Error rates and accuracy on binary class data sets.

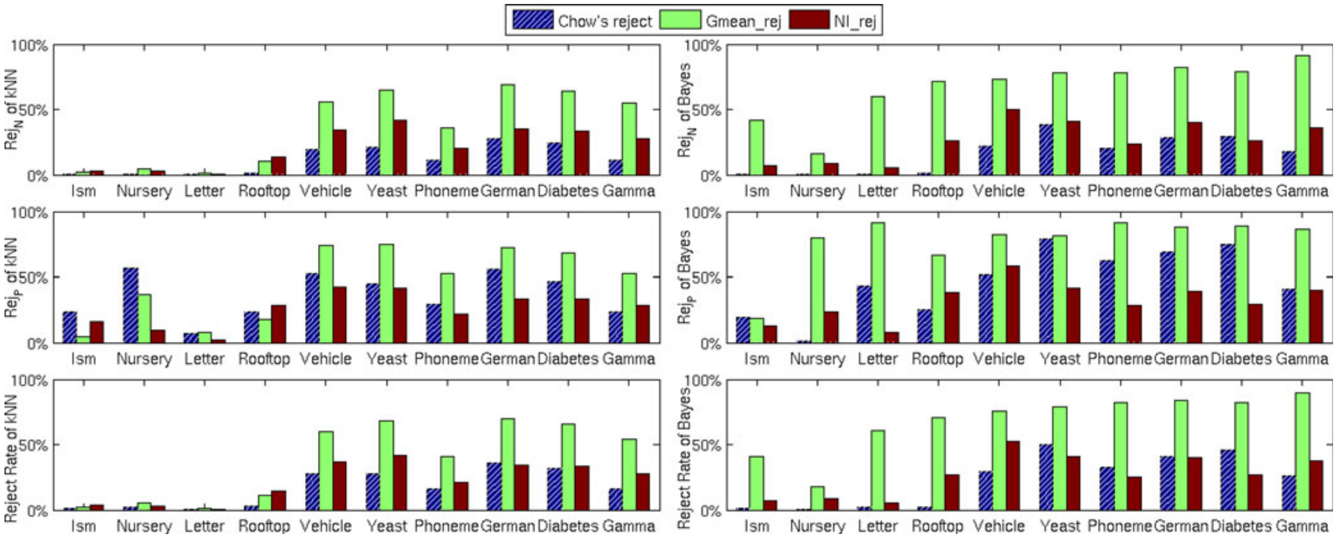


Fig. 4. Reject rates on binary class data sets.

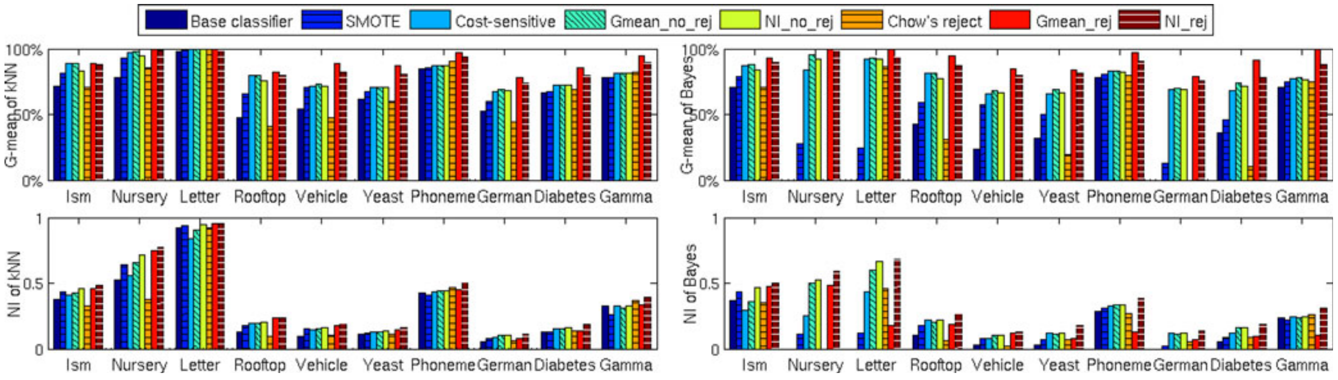


Fig. 5. G-mean and NI on binary class data sets.

for general cost terms. Equation (18a) indicates that the tangent point on the ROC curve will be located at the origin in Fig. 2a, and (18b) demonstrates a graphical interpretation why conventional classifiers fail to process minority class (herein the positive class) properly. However, the situation in (18) can never appear from using the present strategy, because it will result in a zero value of mutual information [41], [45].

Different with the non-abstaining classification, Fig. 2b shows the abstaining classification graphically on an ROC

TABLE 4
The *"Equivalent"* Costs and the Optimal Rejection Thresholds for Binary Class Data Sets

(a) kNN Classifier Based

Data set	$\alpha^*_N\,(\bar{\lambda}_{FP})$	T^*_{rN}	T^*_{rP}	$\bar{\lambda}_{RN}$	$\bar{\lambda}_{RP}$
Ism	0.2312(0.0408)	0.0743(0.0085)	0.7643(0.0296)	0.0272	0.6616
Nursery	0.3482(0.0328)	0.1215(0.0565)	0.7125(0.0403)	0.0288	0.7914
Letter	0.3802(0.0321)	0.1284(0.0343)	0.6140(0.0517)	0.0760	0.4838
Rooftop	0.1372(0.0302)	0.0705(0.0610)	0.7733(0.0404)	0.0544	0.2823
Vehicle	0.1972(0.0769)	0.1101(0.0297)	0.6266(0.0719)	0.1045	0.1556
Yeast	0.3610(0.1333)	0.1245(0.0335)	0.5608(0.0536)	0.0937	0.3414
Phoneme	0.4651(0.0835)	0.1319(0.0180)	0.4543(0.0409)	0.1066	0.2985
German	0.3848(0.0785)	0.1915(0.0336)	0.6021(0.0368)	0.1542	0.3489
Diabetes	0.3725(0.0796)	0.1725(0.0455)	0.5284(0.0672)	0.1585	0.2398
Gamma	0.5682(0.0207)	0.1663(0.0225)	0.4188(0.0319)	0.1376	0.3103

(b) Bayes Classifier Based

Data set	$\alpha^*_N\,(\bar{\lambda}_{FP})$	T^*_{rN}	T^*_{rP}	$\bar{\lambda}_{RN}$	$\bar{\lambda}_{RP}$
Ism	0.1420(0.0230)	0.0222(0.0168)	0.8560(0.0212)	0.0041	0.8198
Nursery	0.1052(0.0117)	0.0593(0.0076)	0.8845(0.0090)	0.0237	0.6242
Letter	0.1155(0.0066)	0.0687(0.0163)	0.8772(0.0128)	0.0273	0.6302
Rooftop	0.0786(0.0299)	0.0242(0.0080)	0.8363(0.0242)	0.0170	0.3147
Vehicle	0.2490(0.0262)	0.1301(0.0149)	0.5369(0.0384)	0.1287	0.1395
Yeast	0.4469(0.0734)	0.2668(0.0147)	0.6413(0.0137)	0.2093	0.4247
Phoneme	0.3364(0.0374)	0.1723(0.0266)	0.5511(0.0363)	0.1641	0.2115
German	0.4265(0.0137)	0.2802(0.0081)	0.6866(0.0085)	0.1736	0.5541
Diabetes	0.3808(0.0460)	0.2461(0.0174)	0.6638(0.0327)	0.2279	0.3020
Gamma	0.5859(0.0483)	0.1723(0.0212)	0.4660(0.0272)	0.1243	0.4028

Optimal values are listed as mean(standard deviation).
(a) Derived based on kNN classifier. (b) Derived based on Bayes classifier.

curve. Two *abstaining slopes*, K_N and K_P, are generally given in the forms of [15]:

$$\begin{aligned} K_N &= \frac{p(N)}{p(P)}\frac{\lambda_{RN}-\lambda_{TN}}{\lambda_{FN}-\lambda_{RP}} = \frac{p(N)}{p(P)}\frac{\overline{\lambda}_{RN}}{1-\overline{\lambda}_{RP}}, \\ K_P &= \frac{p(N)}{p(P)}\frac{\lambda_{FP}-\lambda_{RN}}{\lambda_{RP}-\lambda_{TP}} = \frac{p(N)}{p(P)}\frac{\overline{\lambda}_{FP}-\overline{\lambda}_{RN}}{\overline{\lambda}_{RP}}. \end{aligned} \tag{19}$$

Whenever $K_N \neq K_P$, one can observe the non-zero results of rejection rates. (19) confirms the finding in [20] that at most two independent parameters will determine the rejection range in binary classifications. Sometimes, one can still apply a single independent parameter, such as $K_P = 2K_N$, for abstaining decisions.

TABLE 5
Some ROCCH Vertices of kNN for Diabetes

Index(Label)	ROCCH Vertices (FPR, TPR)	Slope $\widehat{K}$	Threshold τ
16	(0.1188, 0.4883)	2.3156	0.5187
17	(0.1226, 0.4965)	2.1593	0.5106
18(A, C)	**(0.1585, 0.5656)**	**1.9255**	**0.4514**
19	(0.1588, 0.5662)	1.8502	0.4508
20(D)	**(0.1739, 0.5915)**	**1.6778**	**0.4325**
21	(0.1816, 0.6038)	1.5962	0.4211
...	...	...	...
28	(0.3554, 0.7958)	0.8326	0.2617
29(B)	**(0.3615, 0.8002)**	**0.7357**	**0.2586**
30	(0.3634, 0.8016)	0.6921	0.2578
31	(0.3997, 0.8265)	0.6881	0.2334
32	(0.4860, 0.8832)	0.6554	0.1781
33	(0.4901, 0.8858)	0.6543	0.1769
34(E)	**(0.5129, 0.8983)**	**0.5475**	**0.1710**
35	(0.5158, 0.8998)	0.5113	0.1701
36	(0.5389, 0.9104)	0.4605	0.1632
37(F)	**(0.5596, 0.9196)**	**0.4434**	**0.1521**
38	(0.5690, 0.9229)	0.3558	0.1345
39	(0.6098, 0.9349)	0.2928	0.0987

A~F: vertex labels shown in Fig. 6 are bolded
$\widehat{K}$: estimated slope.

There exist relations between rejection thresholds in the *posterior curve plot* [20] and abstaining slopes in the *ROC curve plot*. Their relations and the associated constraint are derived from [20]:

$$K_N = \frac{p(N)}{p(P)}\frac{T_{rN}}{1-T_{rN}}, K_P = \frac{p(N)}{p(P)}\frac{1-T_{rP}}{T_{rP}}, K_N \, < \, K_P \tag{20}$$

The *"equivalent"* costs can be reflected on the ROC curve through the slopes in (16) and (19). Moreover, the optimal rejection thresholds cannot only be associated to the points on the ROC curve directly [15], but also to the slopes in (20). In fact, the slopes in (19) and (20) are equivalent. Based on the performance measures in Fig. 2, the users may adjust the parameters (costs and rejection thresholds) according to their preference both visually and interactively.

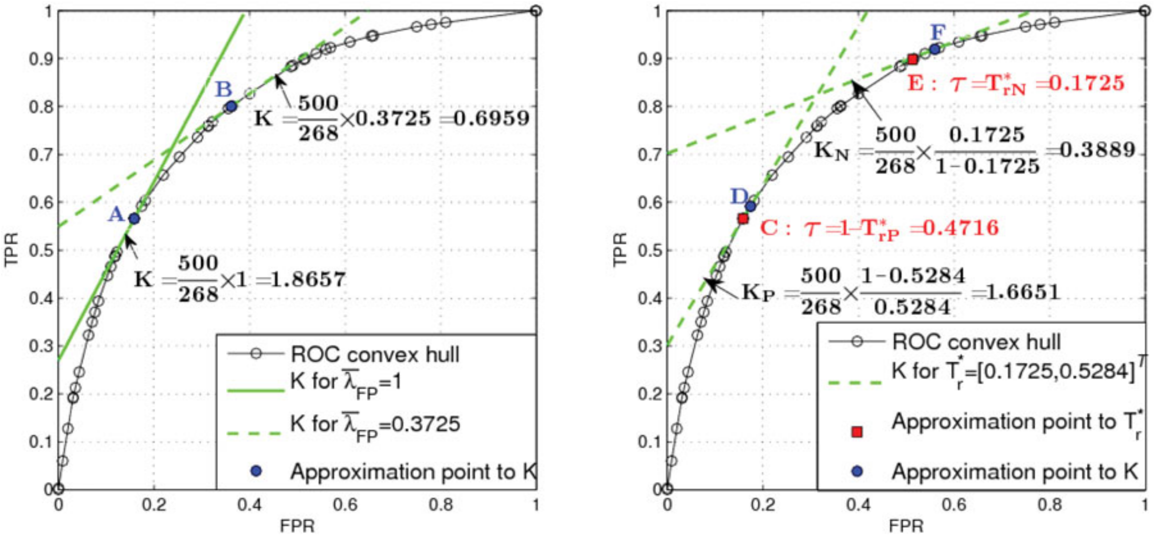
5 Experiments

5.1 Configuration

Table 2 lists 10 binary class and four multi-class data sets with imbalanced class distributions. On *Pageblock*, the maximal class distribution ratio is 175. Most of the data sets are obtained from the UCI Machine Learning Repository,[1] *I sm* is from [25], *Rooftop* is from [18], and *Phoneme* is from KEEL data sets.[2] All of them have continuous attributes and are rescaled to the range $[0, 1]$. The experiments are conducted

[1] http://archive.ics.uci.edu/ml/.

[2] http://sci2s.ugr.es/keel/datasets.php.



(a) For non-abstaining classification

(b) For abstaining classification

Fig. 6. Results on ROCCH of kNN for Diabetes. (a) For non-abstaining classification. (b) For abstaining classification.

using a PC (Intel (R) Core (TM) i7-3770 3.4 GHz, 4 GB memory). The operating system is CentOS 6.4, and the programming language is Matlab. We perform three-fold cross validation and repeat ten times to get average results in all the experiments. All the folds are partitioned in a stratified manner to maintain the class distribution ratio. Table 3 lists the procedure of our NI-based experiments for each run. We adopt nested CV in which the inner CV loop finds the best parameter and the outer CV loop estimates the performance of the model that contains the parameter.

We call our NI-based non-abstaining classification and abstaining classification *"NI_no_rej"* and *"NI_rej"* respectively. To illustrate the effectiveness of our strategy, we adopt *kNN* and *Bayes classifier* as the conventional classifiers. Besides two conventional classifications, we compare our methods with *SMOTE* [25], *Cost-sensitive learning*, *Chow's reject* [10] methods and the G-mean-based methods (*"Gmean_no_rej"* and *"Gmean_rej"*).

In *kNN classifier*, we apply euclidean distance and use the confidence values [42], [49] as the probabilistic outputs. The class assignment is decided by the highest confidence. For brevity, we just list the results of 11-NN on all data sets except 5-NN on *Pageblock*. In *Bayes classifier*, we derive the estimated class-conditional density from the Parzen-window estimation with Gaussian kernel [50] and apply Bayes rule to classification. The smooth parameter is chosen as the average value of the distance from one instance to its rth nearest neighborhood ($r = 10$ empirically), and the empirical probability of the occurrence of class is chosen as the prior probability. In SMOTE, we do not intend to determine the optimal amounts of sampling through any criteria because different criteria may lead to different amounts. Therefore, the average results are presented with the same amount selected from 1 to 5 for all minority classes. In *Cost-sensitive learning*, we simply assign the inverse of the class distribution ratio to the misclassification cost λ_{ij} for $i \neq j$, and $\lambda_{ii} = 0$. We do not consider abstaining for it because the rejection costs would be hard to give. In *Chow's reject*, we simply assign 0.3 to the rejection thresholds for all classes. In G-mean-based methods, we apply our way of parameter settings and optimization to maximize G-mean.

5.2 Evaluation Criterion

In order to show the changes of each class clearly, E_i and Rej_i are applied as the error rate and the reject rate within its ith class respectively. The overall accuracy ("Acc") and the overall reject rate ("Rej") are also applied. *"G"* is short for G-mean with the formula $G{-}mean{=}(\prod_{i=1}^{m} Acc_i)^{\frac{1}{m}}$, where Acc_i represents the accuracy within its ith class.

5.3 Binary Class Tasks

The results on the binary class data sets are shown in Figs. 3, 4 and 5. Both conventional classifiers have low error rates of the negative class and high overall accuracies, but the error rates of the positive class are high. *SMOTE* is an effective method with low error rate of the positive class. However, it does not have the ability to reject instances. *Cost-sensitive learning* performs well under the current cost settings, but its accuracy is the lowest when the class distribution differs greatly. On *Nursery* and *Letter*, the error rates of the positive class are extremely low with *Cost-sensitive learning*, at the price of high error rates of the negative class. Besides, *Gmean_no_rej* and *NI_no_rej* perform well on balancing the classification of two classes. When a reject option is added, the error rate may be reduced and the accuracy may be increased. But it is difficult to decide the rejection costs and the rejection thresholds for lack of information about the rejections. Regarding to *Chow's reject*, it is usually wasteful to reject lots of instances from the positive class with arbitrary settings on the rejection thresholds, while its error rate of the positive class is the highest among the abstaining methods. On most data sets, *Gmean_rej* achieves the highest accuracy and the lowest error rate of the positive class, at the price of considerably high reject rate. However, the accuracy of *Gmean_rej* is lower than the conventional classifications on *Ism* and *Rooftop*. One explanation is that the

TABLE 6
Results on Cardiotocography and Thyroid

(a)

Data set		Method	E_1(%)	E_2(%)	E_3(%)	Acc(%)	Rej_1(%)	Rej_2(%)	Rej_3(%)	Rej(%)	G(%)	NI
Cardiotocography	kNN	kNN classifier	2.39	39.49	28.67	90.29	—	—	—	—	74.88	0.4798
		SMOTE	6.71	23.35	21.26	89.78	—	—	—	—	82.37	0.5346
		Cost-sensitive	100.00	100.00	**0.00**	8.28	—	—	—	—	0.00	0.0000
		Gmean_no_rej	13.20	16.33	15.61	86.17	—	—	—	—	84.86	0.5105
		NI_no_rej	8.59	19.93	19.95	88.90	—	—	—	—	83.56	0.5224
		Chow's reject	0.53	19.66	15.47	94.96	5.70	40.54	27.28	12.32	80.56	0.4566
		Gmean_rej	**0.39**	**4.88**	5.09	**98.10**	19.23	68.77	31.41	27.11	**91.05**	0.4724
		NI_rej	4.44	9.46	15.96	92.94	13.24	25.08	6.34	14.31	88.04	**0.5574**
	Bayes	Bayes classifier	0.06	88.51	76.89	81.31	—	—	—	—	29.44	0.1579
		SMOTE	7.03	41.79	56.52	84.04	—	—	—	—	59.41	0.3378
		Cost-sensitive	100.00	100.00	**0.00**	8.28	—	—	—	—	0.00	0.0000
		Gmean_no_rej	17.28	15.81	22.01	82.53	—	—	—	—	81.49	0.4383
		NI_no_rej	12.13	22.10	27.06	85.25	—	—	—	—	79.13	0.4401
		Chow's reject	**0.00**	40.40	19.63	91.21	4.87	59.60	69.11	17.78	0.00	0.1158
		Gmean_rej	0.08	**0.88**	6.03	**98.16**	60.57	94.03	53.30	64.61	**89.21**	0.2510
		NI_rej	5.59	6.92	30.51	90.59	17.43	33.81	8.18	18.94	81.99	**0.4666**
Thyroid	kNN	kNN classifier	0.08	94.65	52.17	93.88	—	—	—	—	29.11	0.1726
		SMOTE	2.86	79.26	39.69	92.38	—	—	—	—	47.84	0.2245
		Cost-sensitive	100.00	100.00	**0.00**	2.31	—	—	—	—	0.00	0.0000
		Gmean_no_rej	25.69	41.03	22.54	73.59	—	—	—	—	**69.65**	0.2233
		NI_no_rej	5.85	70.92	32.28	90.22	—	—	—	—	56.64	0.2347
		Chow's reject	**0.00**	85.33	30.37	**94.81**	1.05	13.47	35.44	2.47	16.18	0.1277
		Gmean_rej	9.28	**36.47**	21.10	86.61	19.16	28.02	10.61	19.42	68.91	0.2413
		NI_rej	4.68	43.43	22.80	91.55	16.91	29.39	12.75	17.45	63.93	**0.2555**
	Bayes	Bayes classifier	0.06	99.24	87.04	92.86	—	—	—	—	7.24	0.0398
		SMOTE	0.89	95.49	78.46	92.49	—	—	—	—	18.74	0.0708
		Cost-sensitive	100.00	100.00	**0.00**	2.31	—	—	—	—	0.00	0.0000
		Gmean_no_rej	35.04	40.83	25.29	64.89	—	—	—	—	65.67	0.1885
		NI_no_rej	14.40	63.28	34.43	82.64	—	—	—	—	58.12	0.1921
		Chow's reject	**0.03**	97.29	71.09	**93.30**	0.30	1.90	20.35	0.84	8.40	0.0301
		Gmean_rej	2.80	**4.62**	11.29	86.06	77.82	85.94	43.99	77.45	**76.66**	0.1401
		NI_rej	8.67	31.50	29.03	85.80	29.45	42.05	11.42	29.68	60.96	**0.1978**

(b)

Cardiotocography	kNN	α_1^*	α_2^*	α_3^*
		0.2276(0.0592)	0.7138(0.1076)	1
		T_{r1}^*	T_{r2}^*	T_{r3}^*
		0.0953(0.0391)	0.6001(0.1053)	0.7122(0.0888)
	Bayes	α_1^*	α_2^*	α_3^*
		0.1332(0.0265)	0.5071(0.0697)	1
		T_{r1}^*	T_{r2}^*	T_{r3}^*
		0.1530(0.0220)	0.7366(0.0422)	0.8249(0.0362)
Thyroid	kNN	α_1^*	α_2^*	α_3^*
		0.1505(0.0395)	0.7065(0.1291)	1
		T_{r1}^*	T_{r2}^*	T_{r3}^*
		0.0764(0.0269)	0.8058(0.0486)	0.8332(0.0450)
	Bayes	α_1^*	α_2^*	α_3^*
		0.0367(0.0069)	0.4037(0.0751)	1
		T_{r1}^*	T_{r2}^*	T_{r3}^*
		0.0726(0.0171)	0.8996(0.0336)	0.9553(0.0219)

(a) Evaluation of the methods, "–": not available, the best performance in each cell is bolded.
(b) Optimal weights and rejection thresholds are listed as mean(standard deviation).

goal of G-mean-based methods is to maximize the geometric mean of the accuracy within each class rather than the overall accuracy. Compared with *Gmean_rej* and *Chow's reject*, our *NI_rej* performs best on the whole with low error rate of the positive class, high accuracy, a certain amount of reject rate, high G-mean and the highest NI.

Table 4 lists the values of the optimal weight α_N^* and rejection thresholds T_r^*. The last two columns represent the "*equivalent*" rejection costs computed from (14) with the mean values of α_N^* and T_r^*. These values are purely determined by the data sets besides the conventional classification algorithms. They can be adopted as "*objective*" references for CSL while the cost information is unknown. In addition, the "*equivalent*" costs of these data sets are consistent with human assumption, which also reflects the effectiveness of our NI-based strategy.

Due to space limitation, Table 5 just lists some of the vertices from which the *ROC convex hull* (ROCCH) of *kNN* is constructed for *Diabetes*. The vertices on ROCCH (shown in Fig. 6) are generated by threshold averaging [46] from 90 validation sets (ten times stratified three-fold nested CV in the inner CV loops) and are indexed from left to right. Each vertex has its own coordinate values that represent FPR and TPR. The slope of each vertex is approximatively estimated by the secant [15]. Different vertex indicates different threshold. The slope and the threshold decrease from left vertex to right vertex.

In Fig. 6a, according to (16), the solid slope line indicates optimal classifications with equal misclassification costs, while the dash slope line indicates optimal classifications with the "*equivalent*" misclassification cost shown in Table 4a. Depending on the decision method in [15], Vertex *A* and Vertex *B* are the optimal classification decision points chosen by the solid line and the dash line, respectively. Moreover, it is clear to see that *B* has a higher TPR than *A*. In Fig. 6b, according to (20), two dash slope lines indicate optimal classifications with the optimal rejection thresholds. Vertex *D* and Vertex *F* are the optimal

TABLE 7
Results on Car

(a)

Data set		Method	E_1(%)	E_3(%)	E_4(%)	Acc(%)	Rej_1(%)	Rej_3(%)	Rej_4(%)	Rej(%)	G(%)	NI
Car	kNN	kNN classifier	0.50	35.76	62.03	92.18	—	—	—	—	66.65	0.6357
		SMOTE	7.13	29.08	52.41	90.75	—	—	—	—	73.23	0.6554
		Cost-sensitive	100.00	**0.00**	100.00	3.76	—	—	—	—	0.00	0.0000
		Gmean_no_rej	6.13	6.16	7.54	91.45	—	—	—	—	90.62	0.7091
		NI_no_rej	3.31	13.05	24.93	93.83	—	—	—	—	86.28	0.7356
		Chow's reject	**0.00**	1.45	**1.45**	**99.70**	5.70	78.47	97.10	24.37	50.00	0.4574
		Gmean_rej	0.05	4.33	4.06	98.84	11.44	23.64	42.90	21.50	**95.47**	0.6265
		NI_rej	1.94	9.70	20.87	95.34	6.66	0.30	2.46	7.00	88.68	**0.7635**
	Bayes	Bayes classifier	**0.00**	100.00	100.00	70.02	—	—	—	—	0.00	0.0000
		SMOTE	27.51	100.00	100.00	65.91	—	—	—	—	0.00	0.1733
		Cost-sensitive	100.00	**0.00**	100.00	3.76	—	—	—	—	0.00	0.0000
		Gmean_no_rej	26.48	17.07	12.75	75.94	—	—	—	—	80.44	0.4302
		NI_no_rej	26.05	15.40	21.88	76.75	—	—	—	—	79.38	0.4374
		Chow's reject	**0.00**	**0.00**	**0.00**	**98.07**	24.73	100.00	100.00	46.25	0.00	0.1828
		Gmean_rej	10.35	7.72	10.14	87.33	24.02	25.98	44.64	34.06	**81.33**	0.3486
		NI_rej	18.76	14.59	14.64	80.44	10.10	0.14	8.12	11.86	79.64	**0.4388**

(b)

kNN	α_1^*	α_2^*	α_3^*	α_4^*
	0.3029(0.0713)	0.5127(0.1216)	1.0183(0.1751)	1
	T_{r1}^*	T_{r2}^*	T_{r3}^*	T_{r4}^*
	0.2509(0.0514)	0.5751(0.0455)	0.7885(0.0604)	0.8027(0.0394)
Bayes	α_1^*	α_2^*	α_3^*	α_4^*
	0.0997(0.0072)	0.2825(0.0171)	0.9958(0.0984)	1
	T_{r1}^*	T_{r2}^*	T_{r3}^*	T_{r4}^*
	0.3083(0.0379)	0.7365(0.0340)	0.9209(0.0104)	0.9245(0.0115)

(a) Evaluation of the methods, "–": not available, the best performance in each cell is bolded.
(b) Optimal weights and rejection thresholds are listed as mean(standard deviation).

classification decision points chosen by the dash lines. It is reasonable that B lies between D and F, which also reflects the feasibility of our strategy. Meanwhile, the optimal decision thresholds are also marked on their closest vertices C and E, respectively. Theoretically speaking, D and F should be cohere with C and E respectively. Due to approximation, things are not so. In addition, the reference parameters can be adjusted to the preference of the users with the graphical interpretations in Fig. 2.

5.4 Multi-Class Tasks

The detailed results on the multi-class data sets are shown in Tables 6, 7 and 8, including the performance evaluations and the values of the optimal parameters. Both conventional classifiers have low error rates of the majority class, but the error rates of the minority classes are high. *SMOTE* is effective in reducing the error rates of the minority classes except for *car*, because the arbitrarily set amount of sampling is not enough for this data set. *Cost-sensitive learning* classifies all instances to the class that has the minimum number of instances, which is meaningless for classifications. As a result, its accuracy is the lowest and its G-mean and NI are zero. Both *Gmean_no_rej* and *NI_no_rej* perform well on balancing the classification of each class. They have low error rates of the minority classes, high G-mean and high NI. The optimal weight of the majority class derived from *NI_no_rej* is the lowest among all classes. *Chow's reject* and *Gmean_rej* generally reject lots of instances from the minority classes; besides, *Gmean_rej* rejects lots of instances from the majority class. Compared with the above methods, our *NI_rej* performs best on the whole with low error rates of the minority classes, a certain amount of reject rate, high G-mean, and the highest NI.

In summary, the observations above suggest that:

1. Within the CFL category, both *SMOTE* and G-mean-based methods are effective in the class imbalance problems. However, they are unable to handle abstaining classifications.
2. Regarding to *Cost-sensitive learning*, it is feasible to apply the inverses of the class distribution ratios as the misclassification costs on binary class tasks. But on multi-class tasks, it may be ineffective. Moreover, the rejection costs are always hard to get.
3. *Chow's reject* would perform poorly if the rejection thresholds are arbitrarily given.
4. NI-based strategy is a good choice for both non-abstaining and abstaining classifications. It can produce reasonable solutions on the minority classes. Moreover, it can provide reference costs to CSL and reference rejection thresholds to other abstaining classification methods.

5.5 Runtime of Methods

The additional computation cost of our strategy compared with the base classifier is to derive optimal parameters. Table 9 lists the wall-clock time of our methods and the base classifiers. These results are all summarized under the stratified three-fold cross validation. In each fold of deriving optimal parameters, we apply Algorithm 1 10 times according to the procedure in Table 3. Note that the runtime of the methods based on kNN is shorter than that of the methods based on Bayes. This is due to the differences in the runtime and the probabilistic outputs of two base classifiers. In kNN, besides its shorter runtime, its probabilistic outputs of misclassifications are generally closer to the regular decision thresholds, while its probabilistic outputs of right classifications are generally farther from the regular decision thresholds than those in Bayes. All these features make kNN based methods easier to find the weights and rejection thresholds that can handle more misclassifications while causing less impact on right classifications.

TABLE 8
Results on Pageblock

(a)

Data set	Method		E_1(%)	E_3(%)	E_5(%)	Acc(%)	Rej_1(%)	Rej_3(%)	Rej_5(%)	Rej(%)	G(%)	NI
Pageblock	kNN	kNN (k=5)	1.08	42.30	51.03	95.88	—	—	—	—	68.12	0.5466
		SMOTE	2.46	32.65	40.10	95.22	—	—	—	—	74.10	0.5666
		Cost-sensitive	100.00	**0.00**	100.00	0.51	—	—	—	—	0.00	0.0000
		Gmean_no_rej	6.04	9.26	24.48	92.92	—	—	—	—	84.74	0.5899
		NI_no_rej	3.73	15.30	26.66	94.63	—	—	—	—	82.17	0.6019
		Chow's reject	**0.49**	22.74	33.56	**97.40**	2.06	36.37	35.13	4.24	71.33	0.5022
		Gmean_rej	2.13	4.30	**19.84**	96.66	4.00	12.37	19.79	5.32	**86.94**	0.6028
		NI_rej	2.57	11.52	21.53	96.10	3.23	5.89	12.26	3.99	85.52	**0.6179**
	Bayes	Bayes classifier	0.78	56.30	62.86	94.85	—	—	—	—	57.89	0.4559
		SMOTE	2.19	51.61	41.73	94.47	—	—	—	—	66.09	0.4982
		Cost-sensitive	100.00	**0.00**	100.00	0.51	—	—	—	—	0.00	0.0000
		Gmean_no_rej	14.45	17.06	16.23	85.23	—	—	—	—	84.66	0.4900
		NI_no_rej	13.71	17.33	15.47	85.78	—	—	—	—	82.33	0.5055
		Chow's reject	**0.17**	47.85	37.39	**96.91**	2.04	9.93	41.91	5.59	50.21	0.3362
		Gmean_rej	1.26	7.26	**5.59**	96.17	50.29	25.48	60.13	49.46	**87.78**	0.4714
		NI_rej	2.41	11.04	14.71	96.25	11.27	19.26	23.13	12.28	87.15	**0.5749**

(b)

kNN	α_1^*	α_2^*	α_3^*	α_4^*	α_5^*
	0.2234(0.0404)	0.6613(0.2007)	1.0869(0.2546)	0.9581(0.2774)	1
	T_{r1}^*	T_{r2}^*	T_{r3}^*	T_{r4}^*	T_{r5}^*
	0.1049(0.0608)	0.5775(0.1096)	0.8310(0.0667)	0.7229(0.1432)	0.7491(0.1046)

Bayes	α_1^*	α_2^*	α_3^*	α_4^*	α_5^*
	0.0219(0.0055)	0.4162(0.0428)	0.9210(0.2150)	0.7418(0.3106)	1
	T_{r1}^*	T_{r2}^*	T_{r3}^*	T_{r4}^*	T_{r5}^*
	0.0411(0.0125)	0.6620(0.1081)	0.9351(0.0254)	0.5595(0.1233)	0.7622(0.0907)

(a) Evaluation of the methods, "–": not available, the best performance in each cell is bolded.
(b) Optimal weights and rejection thresholds are listed as mean(standard deviation).

TABLE 9
Wall-Clock Time (in Seconds) of Base Classifiers and NI-Based Experiments for Each Run

Data set	kNN Classifier Based			Bayes Classifier Based		
	Base classifier	NI_no_rej	NI_rej	Base classifier	NI_no_rej	NI_rej
Ism	10.0	79.6	54.3	254.1	1637.1	1664.4
Nursery	15.7	120.2	91.7	343.8	2228.7	2258.1
Letter	51.8	354.3	194.0	859.4	5383.8	5460.7
Rooftop	28.1	217.8	123.0	650.6	4200.5	4258.6
Vehicle	0.2	2.4	2.9	4.0	13.4	14.9
Yeast	0.3	4.1	4.9	12.4	39.7	40.8
Phoneme	2.3	21.9	21.9	59.6	389.2	396.3
German	0.3	3.5	4.1	5.7	18.7	20.5
Diabetes	0.1	1.9	2.8	3.3	11.0	12.6
Gamma	34.1	234.1	175.4	742.8	4732.4	4822.5
Cardiotocography	0.9	21.7	11.5	26.6	70.3	45.7
Thyroid	8.2	106.2	51.3	109.1	377.0	285.6
Car	0.4	33.4	11.9	17.0	49.9	31.5
Pageblock	3.2	56.6	30.6	204.3	366.9	312.6

6 Conclusion

In this paper, we propose a new strategy of CFL to deal with the class imbalance problem. Based on the specific property of mutual information that can distinguish different error types and reject types, we seek to maximize it as a general rule for dealing with binary/multi-class classifications with/without abstaining. A unique feature is gained in abstaining classifications when information is unknown about errors and rejects. To our best knowledge, no other existing approach is applicable to this scenario. Moreover, we can derive the "*equivalent*" costs for binary classifications. Generally, the "*equivalent*" costs will be changed accordingly to the distributions of the given data sets. Therefore, the present strategy provides an "*objective*" reference for CSL if users want to adjust the costs. For better understanding ROC curves in binary classifications, graphical interpretations of the theoretical ROC curve plots are explained in terms of the related parameters, such as cost terms and rejection thresholds. Empirical study confirms the advantages of the proposed strategy in solving class imbalance problems. At the same time, we recognize the disadvantage of the work that it will add an extra computational cost over the existing approaches. This difficulty will form a future work for advancing the study.

Another interesting direction for future work is to introduce the new strategy to other learning machines, such as SVM, decision trees, neural networks, etc. The challenge will be a need of specific optimization algorithms for different learning machines.

Acknowledgments

This work was supported in part by NSFC (No. 61075051) and the SPRP-CAS (No. XDA06030300).

References

[1] N. Japkowicz and S. Stephen, "The class imbalance problem: A systematic study," *Intell. Data Anal.*, vol. 6, no. 5, pp. 429–449, Oct. 2002.

[2] T. Raeder, G. Forman, and N. V. Chawla, "Learning from imbalanced data: Evaluation matters," *Data Mining: Foundations and Intelligent Paradigms*, D. E. Holmes and L. C. Jain, eds., New York, NY, USA: Springer, 2012, pp. 315–331.

[3] X. Chai, L. Deng, Q. Yang, and C. X. Ling, "Test-cost sensitive naive Bayes classification," in *Proc. IEEE Int. Conf. Data Mining*, 2004, pp. 51–58.

[4] P. Domingos, "MetaCost: A general method for making classifiers cost-sensitive," in *Proc. ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 1999, pp. 155–164.

[5] B. Zadrozny, J. Langford, and N. Abe, "Cost-sensitive learning by cost-proportionate example weighting," in *Proc. IEEE Int. Conf. Data Mining*, 2003, pp. 435–442.

[6] K. M. Ting, "An instance-weighting method to induce cost-sensitive trees," *IEEE Trans. Knowl. Data Eng.*, vol. 14, no. 13, pp. 659–665, May 2002.

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

[8] V. S. Sheng and C. X. Ling, "Thresholding for making classifiers cost-sensitive," in *Proc. 21st AAAI Nat. Conf. Artif. Intell.*, 2006, pp. 476–481.

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

[10] C. Chow, "On optimum recognition error and reject tradeoff," *IEEE Trans. Inf. Theory*, vol. 16, no. 1, pp. 41–46, Jan. 1970.

[11] T. Pietraszek, "Classification of intrusion detection alerts using abstaining classifiers," *Intell. Data Anal.*, vol. 11, no. 3, pp. 293—316, Aug. 2007.

[12] M.-R. Temanni, S.-A. Nadeem, D. P. Berrar, and J.-D. Zucker, "Aggregating abstaining and delegating classifiers for improving classication performance: An application to lung cancer survival prediction," in *Proc. CAMDA*, 2007.

[13] T. C. Landgrebe, D. M. Tax, P. Paclík, and R. P. Duin, "The interaction between classification and reject performance for distance-based reject-option classifiers," *Pattern Recognit. Lett.*, vol. 27, no. 8, pp. 908–917, Jun. 2006.

[14] C. C. Friedel, U. Rückert, and S. Kramer, "Cost curves for abstaining classifiers," in *Proc. Int. Conf. Mach. Learn. Workshop ROC Anal. Mach. Learn.*, 2006, pp. 33–40.

[15] F. Tortorella, "Reducing the classification cost of support vector classifiers through an ROC-based reject rule," *Pattern Anal. Appl.*, vol. 7, no. 2, pp. 128–143, Jul. 2004.

[16] A. Guerrero-Curieses, J. Cid-Sueiro, R. Alaiz-Rodríguez, and A. R. Figueiras-Vidal, "Local estimation of posterior class probabilities to minimize classification errors," *IEEE Trans. Neural Netw.*, vol. 15, no. 2, pp. 309–317, Mar. 2004.

[17] C. X. Ling and V. S. Sheng, "Cost-sensitive learning and the class imbalance problem," in *Encyclopedia of Machine Learning*. New York, NY, USA: Springer, 2010.

[18] M. A. Maloof, "Learning when data sets are imbalanced and when costs are unequal and unknown," in *Proc. Int. Conf. Mach. Learn. Workshop Learn. Imbalanced Data Sets II*, 2003.

[19] B. Zadrozny and C. Elkan, "Learning and making decisions when costs and probabilities are both unknown," in *Proc. ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2001, pp. 204–213.

[20] B. -G. Hu, "What are the differences between Bayesian classifiers and mutual-information classifiers?" *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 25, no. 2, pp. 249–264, Feb. 2014.

[21] D. J. Hand, "Measuring classifier performance: A coherent alternative to the area under the ROC curve," *Mach. Learn.*, vol. 77, no. 1, pp. 103–123, Oct. 2009.

[22] C. Drummond and R. C. Holte, "Explicitly representing expected cost: An alternative to ROC representation," in *Proc. ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2000, pp. 198–207.

[23] Y. Zhang and Z.-H. Zhou, "Cost-sensitive face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 10, pp. 1758–1769, Oct. 2010.

[24] M. Kubat and S. Matwin, "Addressing the curse of imbalanced training sets: One-sided selection," in *Proc. Int. Conf. Mach. Learn.*, 1997, pp. 179–186.

[25] N. V. Chawla, K. W. Bowyer, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, no. 1, pp. 321–357, Jan. 2002.

[26] M. Lin, K. Tang, and X. Yao, "Dynamic sampling approach to training neural networks for multiclass imbalance classification," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 24, no. 4, pp. 647–660, Apr. 2013.

[27] S. Ertekin, J. Huang, L. Bottou, and L. Giles, "Learning on the border: Active learning in imbalanced data classification," in *Proc. ACM Conf. Inf. Knowl. Manage.*, 2007, pp. 127–136.

[28] Z. Zheng, X. Wu, and R. Srihari, "Feature selection for text categorization on imbalanced data," *ACM SIGKDD Explorations Newslett.*, vol. 6, no. 1, pp. 80–89, Jun. 2004.

[29] M. Wasikowski and X.-W. Chen, "Combating the small sample class imbalance problem using feature selection," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 10, pp. 1388–1400, Oct. 2010.

[30] S. Wang and X. Yao, "Relationships between diversity of classification ensembles and single-class performance measures," *IEEE Trans. Knowl. Data Eng.*, vol. 25, no. 1, pp. 206–219, Jan. 2013.

[31] Y. Park and J. Ghosh, "Ensembles of α-trees for imbalanced classification problems," *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 1, pp. 131–143, Jan. 2014.

[32] D. A. Cieslak, T. R. Hoens, N. V. Chawla, and W. P. Kegelmeyer, "Hellinger distance decision trees are robust and skew-insensitive," *Data Mining Knowl. Discovery*, vol. 24, no. 1, pp. 136–158, Jan. 2012.

[33] L. M. Manevitz and M. Yousef, "One-class SVMs for document classification," *J. Mach. Learn. Res.*, vol. 2, pp. 139–154, 2002.

[34] N. Japkowicz, "Supervised versus unsupervised binary-learning by feedforward neural networks," *Mach. Learn.*, vol. 42, pp. 97–122, Jan. 2001.

[35] Y. Sun, M. S. Kamel, and Y. Wang, "Boosting for learning multiple classes with imbalanced class distribution," in *Proc. IEEE Int. Conf. Data Mining*, 2006, pp. 592–602.

[36] T. Pietraszek, "Optimizing abstaining classifiers using ROC analysis," in *Proc. Int. Conf. Mach. Learn.*, 2005, pp. 665–672.

[37] G. Fumera, F. Roli, and G. Giacinto, "Reject option with multiple thresholds," *Pattern Recognit.*, vol. 33, no. 12, pp. 2099–2101, 2000.

[38] M. Li and I. K. Sethi, "Confidence-based classifier design," *Pattern Recognit.*, vol. 39, no. 7, pp. 1230–1240, Jul. 2006.

[39] B.-G. Hu, R. He, and X.-T. Yuan, "Information-theoretic measures for objective evaluation of classifications," *Acta Automatica Sinica*, vol. 38, no. 7, pp. 1169–1182, Jul. 2012.

[40] J. Principe, D. Xu, Q. Zhao, and J. Fisher, "Learning from examples with information-theoretic criteria," *J. VLSI Signal Process. Syst.*, vol. 26, no. 1/2, pp. 61–77, Aug. 2000.

[41] D. MacKay, *Information Theory, Inference and Learning Algorithms*. Cambridge, U.K.: Cambridge Univ. Press, 2003.

[42] X. Zhang and B.-G. Hu, "Learning in the class imbalance problem when costs are unknown for errors and rejects," in *Proc. IEEE 12th Int. Conf. Data Mining Workshops*, 2012, pp. 194–201.

[43] M. J. D. Powell, "An efficient method for finding the minimum of a function of several variables without calculating derivatives," *Comput. J.*, vol. 7, no. 2, pp. 155–162, 1964.

[44] R. C. Prati, G. E. A. P. A. Batista, and M. C. Monard, "A survey on graphical methods for classification predictive performance evaluation," *IEEE Trans. Knowl. Data Eng.*, vol. 23, no. 11, pp. 1601–1618, Nov. 2011.

[45] B.-G. Hu and Y. Wang, "Evaluation criteria based on mutual information for classifications including rejected class," *Acta Automatica Sinica*, vol. 34, no. 11, pp. 1396–1403, Nov. 2008.

[46] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognit. Lett.*, vol. 27, no. 8, pp. 861–874, Jun. 2006.

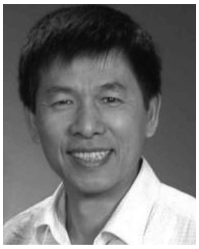
[47] F. Provost and T. Fawcett, "Robust classification for imprecise environments," *Mach. Learn.*, vol. 42, no. 3, pp. 203–231, 2001.

[48] R. O. Duda, P. E. Hart, and D. Stork, *Pattern Classification*. 2nd ed. New York, NY, USA: Wiley, 2003.

[49] J. Arlandis, J. C. Perez-Cortes, and J. Cano, "Rejection strategies and confidence measures for a k-NN classifier in an OCR task," in *Proc. Int. Conf. Pattern Recognit.*, 2002, pp. 576–579.

[50] S. Tong, "Restricted Bayes optimal classifiers," in *Proc. AAAI Nat. Conf. Artif. Intell.*, 2000, pp. 658–664.

Xiaowan Zhang received the BSc degree in automation from Nankai University, China, in 2008. She studied computer science from the National Laboratory of Pattern Recognition (NLPR), Institute of Automation, Chinese Academy of Sciences, Beijing, China. She is currently working toward the PhD degree in computer science at NLPR. Her research interests include machine learning and data mining.



Bao-Gang Hu received the MSc degree from the University of Science and Technology, Beijing, China, in 1983, and the PhD degree from McMaster University, Canada, in 1993, both in mechanical engineering. From 1994 to 1997, he was a research engineer and senior research engineer at C-CORE, Memorial University of Newfoundland, Canada. Currently, he is a professor with the National Laboratory of Pattern Recognition (NLPR), Institute of Automation, Chinese Academy of Science, Beijing, China. From 2000 to 2005, he was the Chinese director of LIAMA (the Chinese-French Joint Laboratory for Computer Science, Control, and Applied Mathematics). His main research interests include pattern recognition and plant growth modeling. He is a senior member of the IEEE.