

Foundation of Mining Class-Imbalanced Data

Da Kuang, Charles X. Ling, and Jun Du

Department of Computer Science
The University of Western Ontario, London, Ontario, Canada N6A 5B7
{dkuang, cling, jdu42}@csd.uwo.ca

Abstract. Mining class-imbalanced data is a common yet challenging problem in data mining and machine learning. When the class is imbalanced, the error rate of the rare class is usually much higher than that of the majority class. How many samples do we need in order to bound the error of the rare class (and the majority class)? If the misclassification cost of the class is known, can the cost-weighted error be bounded as well? In this paper, we attempt to answer those questions with PAC-learning. We derive several upper bounds on the sample size that guarantee the error on a particular class (the rare and majority class) and the cost-weighted error, with the consistent and agnostic learners. Similar to the upper bounds in traditional PAC learning, our upper bounds are quite loose. In order to make them more practical, we empirically study the pattern observed in our upper bounds. From the empirical results we obtain some interesting implications for data mining in real-world applications. As far as we know, this is the first work providing theoretical bounds and the corresponding practical implications for mining class-imbalanced data with unequal cost.

1 Introduction

In data mining, datasets are often imbalanced (or class imbalanced); that is, the number of examples of one class (the rare class) is much smaller than the number of the other class (the majority class).¹

This problem happens often in real-world applications of data mining. For example, in medical diagnosis of a certain type of cancer, usually only a small number of people being diagnosed actually have the cancer; the rest do not. If the cancer is regarded as the positive class, and non-cancer (healthy) as negative, then the positive examples may only occur 5% in the whole dataset collected. Besides, the number of fraudulent actions is much smaller than that of normal transactions in credit card usage data. When a classifier is trained on such an imbalanced dataset, it often shows a strong bias toward the majority class, since the goal of many standard learning algorithms is to minimize the overall prediction error rate. Thus, by simply predicting every example as the majority class, the classifier can still achieve a very low error rate on a class-imbalanced dataset with, for example, 2% rare class.

When mining the class-imbalanced data, do we always get poor performance (e.g., 100% error) on the rare class? Can the error of the rare class (as well as the majority class) be bounded? If so, is the bound sensitive to the class imbalance ratio? Although

¹ In this paper, we only study binary classification.

the issue of class imbalance has been received extensive studies [9, 3, 2, 7, 4, 5], as far as we know, no previous works have been done to answer those questions.

In fact, PAC learning (Probably Approximately Correct Learning) [8, 6] is an appropriate model to study the bounds for classification performance. The traditional PAC learning model studies the learnability of the general concept for certain kinds of learners (such as consistent learner and agnostic learner), and answers the question that how many examples would be sufficient to guarantee a low total error rate. However, previous works [9] point out that accuracy or total error rate are inappropriate to evaluate the classification performance when class is imbalanced, since such metrics overly emphasize the majority class and neglect the rare class which is usually more important in real-world applications. Thus, when class is imbalanced, better measures are desired. In our paper, we will use error rate on the rare (and majority) class and *cost-weighted error*² to evaluate the classification performance on class-imbalanced data. The error rate on the rare (and majority) class can reflect how well the rare (and majority) class is learned. If the misclassification cost of the class is known, we can adopt another common measure (cost-weighted error) to deal with imbalanced data. By weighting the error rate on each class by its associated cost, we will get higher penalty for the error on the rare class (usually the more important class).

In our paper, we attempt to use the PAC-learning model to study, when class is imbalanced, how many sampled examples needed to guarantee a low error on a particular class (the rare class or majority class) and a low cost-weighted error respectively. A bound on cost-weighted error is necessary since it would naturally “suppress” errors on the rare class. We theoretically derive several upper bounds for both consistent learner and agnostic learner. Similar to the upper bounds in traditional PAC learning, the bounds we derive are generally quite loose, but they do provide a theoretical guarantee on the classification performance when class-imbalanced data is learned. Due to the loose bounds, to make our work more practical, we also empirically study how class imbalance affects the performance by using a specific learner. From our experimental results, some interesting implications can be found. The results in this paper can provide some theoretical foundations for mining the class-imbalanced data in the real world.

The rest of the paper is organized as follows. We theoretically derive several upper bounds on the sample complexity for both consistent learner and agnostic learner. Then we empirically explore how class imbalance affects the classification performance by using a specific learner. Finally, we draw the conclusions and address our future work.

2 Upper Bounds

In this section, we take advantage of PAC-learning theory to study the sample complexity when learning from the class-imbalanced data. Instead of bounding the total error rate, we focus on the error rate on a particular class (rare class or majority class) and the cost-weighted error.

² We will define it in the next section.

2.1 Error Rate on a Particular Class

First of all, we introduce some notations for readers' convenience. We assume that the examples in training set T are drawn randomly and independently from a fixed but unknown class-imbalanced distribution D . We denote p ($0 < p < 0.5$) as the proportion of the rare class (the positive class) in D . For the class-imbalanced training set, p can be very small (such as 0.001). The number of total training examples is denoted as m and the number of positive and negative training examples are denoted as m_+ and m_- respectively. For any hypothesis h from the hypothesis space H , we denote $e_D(h)$, $e_{D_+}(h)$, and $e_{D_-}(h)$ as the total, the positive, and the negative generalization error, respectively, of h , and we also denote $e_T(h)$, $e_{T_+}(h)$, and $e_{T_-}(h)$ as the total, the positive, and the negative training error, respectively, of h .

Given ε ($0 < \varepsilon < 1$) and δ ($0 < \delta < 1$), the traditional PAC learning provides upper bounds on the total number of training examples needed to guarantee $e_D(h) < \varepsilon$ with probability at least $1 - \delta$. However, it guarantees nothing about the positive error $e_{D_+}(h)$ for the imbalanced datasets. As we discussed before, the majority classifier would predict every example as negative, resulting in a 100% error rate on the positive (rare) examples. To have a lower positive error, the learner should observe more positive examples. Thus, in this subsection, we study the upper bounds of the examples on a particular class (say positive class here) needed to guarantee, with probability at least $1 - \delta$, $e_{D_+}(h) < \varepsilon_+$, given any ε_+ ($0 < \varepsilon_+ < 1$).

We first present a simple relation between the total error and the positive error as well as the negative error, and will use it to derive some upper bounds.

Theorem 1. *Given any ε_+ ($0 < \varepsilon_+ < 1$) and the positive class proportion p ($0 < p < 0.5$) according to distribution D and target function C , for any hypothesis h , if $e_D(h) < \varepsilon_+ \times p$, then $e_{D_+}(h) < \varepsilon_+$.*

Proof. To prove this, we simply observe that,

$$e_D(h) = e_{D_+}(h) \times p + e_{D_-}(h) \times (1 - p) \geq e_{D_+}(h) \times p.$$

Thus,

$$e_{D_+}(h) \leq \frac{e_D(h)}{p}.$$

Therefore, if $e_D(h) < \varepsilon_+ \times p$, $e_{D_+}(h) < \varepsilon_+$.

Following the same direction, we can also derive a similar result for the error on negative class $e_{D_-}(h)$. That is, given ε_- ($0 < \varepsilon_- < 1$), if $e_D(h) < \varepsilon_- \times (1 - p)$, then $e_{D_-}(h) < \varepsilon_-$.

Theorem 1 simply tells us, as long as the total error is small enough, a desired positive error (as well as negative error) can always be guaranteed. Based on Theorem 1, we can "reuse" the upper bounds in the traditional PAC learning model and adapt them to be the upper bounds of a particular class in the class-imbalanced datasets. We first consider consistent learner in the next subsection.

Consistent Learner We consider *consistent learner* L using hypothesis space H by assuming that the target concept c is representable by H ($c \in H$). Consistent learner always makes correct prediction on the training examples. Let us assume that $UB(\varepsilon, \delta)$ is an upper bound on the sample size in the traditional PAC-learning, which means that, given ε ($0 < \varepsilon < 1$) and δ ($0 < \delta < 1$), if the total number of training examples $m \geq UB(\varepsilon, \delta)$, a consistent learner will produce a hypothesis h such that with the probability at least $(1 - \delta)$, $e_D(h) \leq \varepsilon$. The following theorem shows that we can adapt any upper bound in the traditional PAC-learning to the bounds that guarantee a low error on the positive class and negative class respectively.

For any upper bound of a consistent PAC learner $UB(\varepsilon, \delta)$, we can always replace ε in $UB(\varepsilon, \delta)$ with $\varepsilon_+ \times p$ or $\varepsilon_- \times (1 - p)$, and consequently obtain a upper bound to guarantee the error rate on that particular class.

Theorem 2. *Given $0 < \varepsilon_+ < 1$, if the number of positive examples*

$$m_+ \geq UB(\varepsilon_+ \times p, \delta) \times p,$$

then with probability at least $1 - \delta$, the consistent learner will output a hypothesis h having $e_{D^+}(h) \leq \varepsilon_+$.

Proof. By the definition of the upper bound for the sample complexity, given $0 < \varepsilon < 1$, $0 < \delta < 1$, if $m \geq UB(\varepsilon, \delta)$, with probability at least $1 - \delta$ any consistent learner will output a hypothesis h having $e_D(h) \leq \varepsilon$.

Here, we simply substitute ε in $UB(\varepsilon, \delta)$ with $\varepsilon_+ \times p$, which is still within $(0, 1)$. Consequently, we obtain that if $m \geq UB(\varepsilon_+ \times p, \delta)$, with probability at least $1 - \delta$ any consistent learner will output a hypothesis h having $e_D(h) \leq \varepsilon_+ \times p$. According to Theorem 1, we get $e_{D^+}(h) < \varepsilon_+$.

Also, $m = \frac{m_+}{p}$, thus we know, $m \geq UB(\varepsilon_+ \times p, \delta)$ equals to

$$m_+ \geq UB(\varepsilon_+ \times p, \delta) \times p.$$

Thus, the theorem is proved.

By using the similar proof to Theorem 2, we can also derive the upper bound for the negative class. Given $0 < \varepsilon_- < 1$, if the number of negative examples $m_- \geq UB(\varepsilon_- \times (1 - p), \delta) \times (1 - p)$, then, with probability at least $1 - \delta$, the consistent learner will output a hypothesis h having $e_{D^-}(h) \leq \varepsilon_-$.

The two upper bounds above can be adapted to any traditional upper bound of consistent learners. For instance, it is well known that any consistent learner using finite hypothesis space H has an upper bound $\frac{1}{\varepsilon} \times (\ln|H| + \ln\frac{1}{\delta})$ [6]. Thus, by applying our new upper bounds, we obtain the following corollary.

Corollary 1. *For any consistent learner using finite hypothesis space H , the upper bound on the number of positive sample for $e_{D^+}(h) \leq \varepsilon_+$ is*

$$m_+ \geq \frac{1}{\varepsilon_+} (\ln|H| + \ln\frac{1}{\delta}),$$

and the upper bound on the number of negative sample for $e_{D_-}(h) \leq \varepsilon_-$ is

$$m_- \geq \frac{1}{\varepsilon_-} (\ln|H| + \ln \frac{1}{\delta}).$$

From Corollary 1, we can discover that when the consistent learner uses *finite* hypothesis space, the upper bound of sample size on a particular class is directly related to the desired error rate (ε_+ or ε_-) on the class, and the class imbalance ratio p does not affect the upper bound. This indicates that, for consistent learner, no matter how class-imbalanced the data is (how small p is), as soon as we sample sufficient examples in a class, we can always achieve the desired error rate on that class.

Agnostic Learner In this subsection, we consider *agnostic learner* L using finite hypothesis space H , which makes *no* assumption about whether or not the target concept c is representable by H . Agnostic learner simply finds the hypothesis with the minimum (probably non-zero) training error. Given an arbitrary small ε_+ , we can not ensure $e_{D_+}(h) \leq \varepsilon_+$, since very likely $e_{T_+}(h) > \varepsilon_+$. Hence, we guarantee $e_{D_+}(h) \leq e_{T_+}(h) + \varepsilon$ to happen with probability higher than $1 - \delta$, for such h with the minimum training error. To prove the upper bound for agnostic learner, we adapt the original proof for agnostic learner in [6]. The original proof regards drawing m examples from the distribution D as m independent Bernoulli trials, but in our proof, we only treat drawing m_+ examples from the positive class as m_+ Bernoulli trials.

Theorem 3. Given ε_+ ($0 < \varepsilon_+ < 1$), any δ ($0 < \delta < 1$), if the number of positive examples observed

$$m_+ > \frac{1}{2\varepsilon_+^2} (\ln|H| + \ln \frac{1}{\delta}),$$

then with probability at least $1 - \delta$, the agnostic learner will output a hypothesis h , such that $e_{D_+}(h) \leq e_{T_+}(h) + \varepsilon_+$

Proof. For any h , we consider $e_{D_+}(h)$ as the true probability that h will misclassify a randomly drawn positive example. $e_{T_+}(h)$ is an observed frequency of misclassification over the given m_+ positive training examples. Since the entire training examples are drawn identically and independently, drawing and predicting positive training examples are also identical and independent. Thus, we can treat drawing and predicting m_+ positive training examples as m_+ independent Bernoulli trials.

Therefore, according to Hoeffding bounds, we can have,

$$Pr[e_{D_+}(h) > e_{T_+}(h) + \varepsilon] \leq e^{-2m_+\varepsilon^2}.$$

According to the inequation above, we can derive,

$$Pr[(\exists h \in H)(e_{D_+}(h) > e_{T_+}(h) + \varepsilon)] \leq |H|e^{-2m_+\varepsilon^2}.$$

This formula tells us that the probability that there exists one bad hypothesis h making $e_{D_+}(h) > e_{T_+}(h) + \varepsilon$ is bounded by $|H|e^{-2m_+\varepsilon^2}$. If we let $|H|e^{-2m_+\varepsilon^2}$ be less than δ ,

then for any hypothesis including the outputted hypothesis h in H , $e_{D_+}(h) - e_{T_+}(h) \leq \varepsilon$ will hold true with the probability at least $1 - \delta$. So, solving for m_+ in the inequation $|H|e^{-2m_+\varepsilon^2} < \delta$, we obtain

$$m_+ > \frac{1}{2\varepsilon_+^2} (\ln|H| + \ln\frac{1}{\delta}).$$

Thus, the theorem is proved.

In fact, by using the similar procedure, we can also prove the upper bound for the number of negative examples m_- when using agnostic learner: $\frac{1}{2\varepsilon_-^2} (\ln|H| + \ln\frac{1}{\delta})$.

We can observe a similar pattern here. The upper bounds for the agnostic learner are also not affected by the class imbalance ratio p .

From the upper bound of either consistent learner or agnostic learner we derived, we learned that when the amount of examples on a class is enough, class imbalance does not take any effect. This discovery actually refutes a common misconception that we need more examples just because of the more imbalanced class ratio. We can see, the class imbalance is in fact a data insufficiency problem, which was also observed empirically in [4]. Here, we further confirm it with our theoretical analysis.

In this subsection, we derive a new relation (Theorem 1) between the positive error and the total error, and use it to derive a general upper bound (Theorem 2) which can be applied to any traditional PAC upper bound for consistent learner. We also extend the existing proof of agnostic learner to derive a upper bound on a particular class for agnostic learner. Although the proof of the theorems above may seem straightforward, no previous work explicitly states the same conclusion from the theoretical perspective.

It should be noted that although the agnostic learner outputs the hypothesis with the minimum (total) training error, it is possible that the outputted hypothesis has 100% error rate on the positive class in the training set. In this case, the guaranteed small difference ε_+ between the true positive error and the training positive error can still result in 100% true error rate on the positive class. If the positive errors are more costly than the negative errors, it is more reasonable to assign higher cost for misclassifying positive examples, and let the agnostic learner minimize the cost-weighted training error instead of the flat training error. In the following part, we will introduce misclassification cost to our error bounds.

2.2 Cost-Weighted Error

In this subsection, we take misclassification cost into consideration. We assume that the misclassification cost of the class is known, and the cost of a positive error (rare class) is higher than (at least equals) the cost of a negative error. We use C_{FN} and C_{FP} to represent the cost of misclassifying a positive example and a negative example, respectively.³ And we denote r as the cost ratio, $\frac{C_{FN}}{C_{FP}}$ ($r \geq 1$). Here we define a new type of error, named *cost-weighted error*.

³ We assume the cost of correctly predicting a positive example and a negative example is 0, meaning that $C_{TP} = 0$ and $C_{TN} = 0$.

Definition 1 (Cost-Weighted Error). Given the cost ratio r , the class ratio p , e_{D+} as the positive error on D , e_{D-} as the negative error on D , the cost-weighted error on D can be defined as,

$$c_D(h) = \frac{rpe_{D+} + (1-p)e_{D-}}{rp + (1-p)}.$$

By the same definition, we can also define the cost-weighted error on the training set T as $c_T(h) = \frac{rpe_{T+} + (1-p)e_{T-}}{rp + (1-p)}$. The weight of the error on a class is determined by its class ratio and misclassification cost. The rp is the weight for the positive class and $1-p$ is the weight for the negative class. In our definition for the cost-weighted error, we use the normalized weight.

In the following part, we study the upper bounds for the examples needed to guarantee a low cost-weighted error on D . We give a non-trivial proof for the upper bounds of consistent learner, and the proof for the upper bound of agnostic learner is omitted due to its similarity to that of the consistent learner (but only with finite hypothesis space).

Consistent Learner To derive a relatively tight upper bound of sample size for cost-weighted error, we first introduce a property. That is, there exist many combinations of positive error e_{D+} and negative error e_{D-} that can make the same cost-weighted error value. For example, given $rp = 0.4$, if $e_{D+} = 0.1$ and $e_{D-} = 0.2$, c_D will be 0.16, while $e_{D+} = 0.25$ and $e_{D-} = 0.1$ can also produce the same cost-weighted error. We can let the upper bound to be the least required sample size among all the combinations of positive error and negative error that can make the desired cost-weighted error.

Theorem 4. Given ε ($0 < \varepsilon < 1$), any δ ($0 < \delta < 1$), the cost ratio r ($r \geq 1$) and the positive proportion p ($0 < p < 0.5$) according to the distribution D , if the total number of examples observed

$$m \geq \frac{1+r}{\varepsilon(rp + (1-p))} (\ln|H| + \ln \frac{1}{\delta}),$$

then, with probability at least $1 - \delta$, the consistent learner will output a hypothesis h such that the cost-weighted error $c_D(h) \leq \varepsilon$.

Proof. In order to make $c_D(h) \leq \varepsilon$, we should ensure,

$$\frac{rpe_{D+} + (1-p)e_{D-}}{rp + (1-p)} \leq \varepsilon. \quad (1)$$

Here, we let $X = \frac{rp}{rp + (1-p)}$, thus $1 - X = \frac{(1-p)}{rp + (1-p)}$. Accordingly, Formula (1) can be transformed into $Xe_{D+} + (1-X)e_{D-} \leq \varepsilon$. To guarantee it, we should make sure,

$$e_{D-} \leq \frac{\varepsilon - Xe_{D+}}{1-X}.$$

According to Corollary 1, if we observe,

$$m_- \geq \frac{1}{\frac{\varepsilon - Xe_{D+}}{1-X}} (\ln|H| + \ln \frac{1}{\delta}), \quad (2)$$

we can also ensure $e_{D_-}(h) \leq \frac{\varepsilon - X e_{D_+}}{1-X}$ with probability at least $1 - \delta$ to happen. Besides, in order to have e_{D_+} on positive class, we also need to observe,

$$m_+ \geq \frac{1}{e_{D_+}} (\ln|H| + \ln \frac{1}{\delta}). \quad (3)$$

To guarantee Formula (2) and (3), we need to sample at least m examples such that $m = \text{MAX}(\frac{m_+}{p}, \frac{m_-}{1-p})$. Thus,

$$m \geq \text{MAX}\left(\frac{1}{e_{D_+} \times p}, \frac{1}{\frac{\varepsilon - X e_{D_+}}{1-X} \times (1-p)}\right) (\ln|H| + \ln \frac{1}{\delta}).$$

However, since e_{D_+} is a variable, different e_{D_+} will lead to different e_{D_-} , and thus affect m . In order to have a tight upper bound for m , we only need,

$$m \geq \text{MIN}_{0 \leq e_{D_+} \leq \frac{\varepsilon}{X}} \left(\text{MAX}\left(\frac{1}{e_{D_+} \times p}, \frac{1}{\frac{\varepsilon - X e_{D_+}}{1-X} \times (1-p)}\right) (\ln|H| + \ln \frac{1}{\delta}) \right).$$

When $\frac{1}{e_{D_+} \times p} > \frac{1}{\frac{\varepsilon - X e_{D_+}}{1-X} \times (1-p)}$, $\text{MAX}\left(\frac{1}{e_{D_+} \times p}, \frac{1}{\frac{\varepsilon - X e_{D_+}}{1-X} \times (1-p)}\right) = \frac{1}{e_{D_+} \times p}$, which is a decreasing function of e_{D_+} , but when $\frac{1}{e_{D_+} \times p} < \frac{1}{\frac{\varepsilon - X e_{D_+}}{1-X} \times (1-p)}$, it becomes an increasing function of e_{D_+} . Thus, the minimum value of the function can be achieved when $\frac{1}{e_{D_+} \times p} = \frac{1}{\frac{\varepsilon - X e_{D_+}}{1-X} \times (1-p)}$. By solving the equation, we obtain the minimum value for the function,

$$\frac{1}{\frac{\varepsilon(1-p)}{p+X-2Xp} \times p} (\ln|H| + \ln \frac{1}{\delta}).$$

If we recover X with $\frac{rp}{rp+(1-p)}$, then it can be transformed into $\frac{1+r}{\varepsilon(rp+(1-p))} (\ln|H| + \ln \frac{1}{\delta})$. Therefore, as long as,

$$m \geq \frac{1+r}{\varepsilon(rp+(1-p))} (\ln|H| + \ln \frac{1}{\delta}),$$

then with probability at least $1 - \delta$, the consistent learner will output a hypothesis h such that $c_D(h) \leq \varepsilon$.

We can see that the upper bound of cost-weighted error for consistent learner is related to p and r . By performing a simple transformation, we can transform the above upper bound into $\frac{r+1}{\varepsilon((r-1)p+1)} (\ln|H| + \ln \frac{1}{\delta})$. It is known that $r \geq 1$, thus $r-1 \geq 0$. Therefore, as p decreases within $(0, 0.5)$, the upper bound increases. It means that the more the class is imbalanced, the more examples we need to achieve a desired cost-weighted error. In this case, class imbalance actually affects the classification performance in terms of cost-weighted error. If we make another transformation to the upper bound, we can obtain, $\frac{1}{p\varepsilon} + \frac{2p-1}{\varepsilon(rp^2+(1-p)p)} (\ln|H| + \ln \frac{1}{\delta})$. Since $0 < p < 0.5$, $2p-1 < 0$. Thus, as r increases, the upper bound also increases. It shows that a higher cost ratio $\frac{C_{FN}}{C_{FP}}$ would require more examples for training. Intuitively speaking, when class is imbalanced, the

cost-weighted error largely depends on the error on the rare class. As we have proved before, to achieve the same error on the rare class, we need the same amount of examples on the rare class, thus more class-imbalanced data requires more examples in total. Besides, higher cost on the rare class leads to higher cost-weighted error, thus to achieve the same cost-weighted error, we will also need more examples in total.

Agnostic Learner As mentioned before, the hypothesis with the minimum training error produced by agnostic learner may still lead to 100% error rate on the rare class. Hence, instead of outputting the hypothesis with minimum training error, we redefine agnostic learner as the learner that outputs the hypothesis with the minimum cost-weighted error on the training set. Generally, with higher cost on positive errors, the agnostic learner is less likely to produce a hypothesis that misclassifies all the positive training examples. The following theorem demonstrates that, for agnostic learner, how many examples needed to guarantee a small difference of the cost-weighted errors between the distribution D and the training set T .

Theorem 5. *Given ε ($0 < \varepsilon < 1$), any δ ($0 < \delta < 1$), the cost ratio r ($r \geq 1$) and the positive proportion p ($0 < p < 0.5$) according to the distribution D , if the total number of examples observed*

$$m \geq \frac{r\sqrt{p} + \sqrt{1-p}}{2\varepsilon^2(rp + (1-p))} (\ln|H| + \ln\frac{1}{\delta}),$$

then, with probability at least $1 - \delta$, the agnostic learner will output a hypothesis h such that $c_D(h) \leq c_T(h) + \varepsilon$.

The proof for Theorem 5 is very similar to that of Theorem 4, thus here we omit the detail of the proof. Furthermore, we can also extract the same patterns from the upper bound here as found for the upper bound in Theorem 4: more examples are required when the cost ratio increases or the class becomes more imbalanced.

To summarize, in this section we derive several upper bounds to guarantee the error rate on a particular class (rare class or majority class) as well as the cost-weighted error, for both consistent learner and agnostic learner. We found some interesting and useful patterns from those theoretical results: the upper bound for the error rate on a particular class is not affected by the class imbalance, while the upper bound for the cost-weighted error is sensitive to both the class imbalance and the cost ratio. Although those pattern may not be so surprising, as far as we know, no previous work theoretically proved it before. Such theoretical results would be more reliable than the results only based on the empirical observation.

Since the upper bounds we derive are closely related to the hypothesis space, which is often huge for many learning algorithms, they are generally very loose (It should be noted that in traditional PAC learning, the upper bounds are also very loose). In fact, when we practically use some specific learners, to achieve a desired error rate on a class or cost-weighted cost, usually the number of examples needed are much less than the theoretical upper bounds. Therefore, in the next section, we will empirically study the performance of a specific learner, to see how the class imbalance and cost ratio influence the classification performance.

3 Empirical Results with Specific Learner

In this section, we empirically explore the patterns found in the theoretical upper bounds. We hope to see, in practice, how the class imbalance and cost ratio affect the actual examples needed and whether the empirical results reflect our theories. Those empirical observations can be useful for practical data mining or machine learning with class-imbalanced data. In our following experiments, we will empirically study the performance of unpruned decision tree (consistent learner) on class-imbalanced datasets.⁴

3.1 Datasets and Settings

We choose unpruned decision tree for our empirical study, since it is a consistent learner in any case. It can be always consistent with the training data of any concept by building up a full large tree, if there are no conflicting examples with the same attribute values but different class labels. For the specific implementation, we use WEKA [10] and select *J48* with the pruning turned off and the parameter *MinNumObj* = 1.

We create one artificial dataset and select two real-world datasets. The artificial dataset we use is generated by a tree function with five relevant attributes, $A1 - A5$, and six leaves, as shown in Figure 1. To simulate the real-world dataset, we add another 11 irrelevant attributes. Therefore, with 16 binary attributes, we can generate $2^{16} = 65,536$ different examples, and label them with the target concept (28,672 positive and 36,864 negative). We also choose two UCI [1] real-world datasets (*Chess* and *Splice*). In order to make the unpruned decision tree with all the training examples, the conflicting examples (i.e., the examples with identical attribute values but different labels) are eliminated during the pre-process.

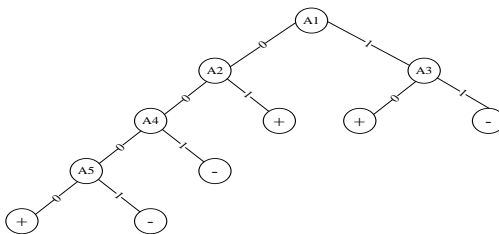


Fig. 1: Artificial tree function.

3.2 Experimental Design and Results

To see how class imbalance affects the error rate on a particular class (here we choose positive class), we compare the positive error under different class ratios but with the same number of positive examples in the training set.

⁴ Due to the limited pages, we only empirically study the consistent learner here.

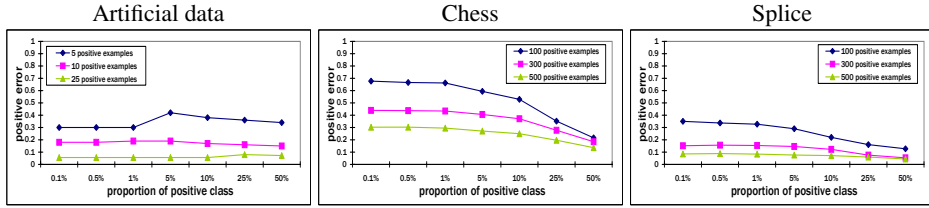


Fig. 2: Positive error of unpruned decision tree on three datasets.

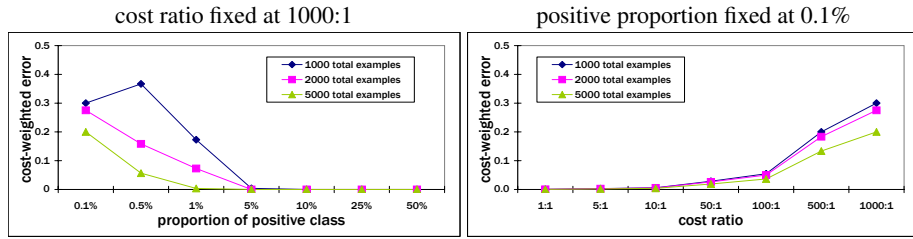


Fig. 3: Cost-weighted error of unpruned decision tree on the artificial data.

Specifically, we manually generate different data distributions with various class ratios where training set and test set are drawn. For example, to generate a data distribution with 10% positive proportion, we simply set the probability of drawing a positive example to be $\frac{1}{9}$ of the probability of drawing a negative example, and the probability of drawing examples within a class is uniform. According to the data distribution, we sample a training set until it contains a certain number of positive examples (we set three different numbers for each dataset), and train a unpruned decision tree on it. Then, we evaluate its performance (positive error and cost-weighted error) on another sampled test set from the same data distribution. Finally, we compare the performance under different data distributions (0.1%, 0.5%, 1%, 5%, 10%, 25%, 50%) to see how class imbalance ratio affects the performance of the unpruned decision tree. All the results are the average value over 10 independent runs.

Figure 2 presents the positive error on three datasets. The three curves in each subgraph represent three different numbers of positive examples in the training set. For the artificial dataset, since the concept is easy to learn, the number of positive examples chosen is smaller than that of the UCI datasets. We can see, generally the more the positive examples for training, the flatter the curve and the lower the positive error. It means, as we have more positive examples, class imbalance has less negative effect on the positive error in practice. The observation is actually consistent with Corollary 1.

To see how class imbalance influences the cost-weighted error, we compare the cost-weighted error under different class ratios with fixed cost ratio. To explore how cost ratio affects the cost-weighted error, we compare the cost-weighted error over different cost ratios with fixed class ratio. For this part, we only use the artificial dataset to show the results (see Figure 3). We can see, generally, as the class becomes more imbalanced

or the cost ratio increases, the cost-weighted error goes higher. It is also consistent with our theory (Theorem 4).

We have to point out that, our experiment is not a verification of our derived theories. The actual amount of examples we used in our experiment is much smaller compared to the theoretical bounds. Despite of that, we still find that the empirical observations have similar patterns to our theoretical results. Thus, our theorems not only offer a theoretical guarantee, but also has some useful implications for real-world applications.

4 Conclusions

In this paper, we study the class imbalance issue from PAC-learning perspective. An important contribution of our work is that, we theoretically prove that the upper bound of the error rate on a particular class is not affected by the (imbalanced) class ratio. It actually refutes a common misconception that we need more examples just because of the more imbalanced class ratio. Besides the theoretical theorems, we also empirically explore the issue of the class imbalance. The empirical observations reflect the patterns we found in our theoretical upper bounds, which means our theories are still helpful for the practical study of class-imbalanced data.

Although intuitively our results might seem to be straightforward, few previous works have explicitly addressed these fundamental issues with PAC bounds for class-imbalanced data before. Our work actually confirms the practical intuition by theoretical proof and fills a gap in the established PAC learning theory. For imbalanced data issue, we do need such a theoretical guideline for practical study.

In our future work, we will study bounds for AUC, since it is another useful measure for the imbalanced data. Another common heuristic method to deal with imbalanced data is over-sampling and under-sampling. We will also study their bounds in the future.

References

1. Asuncion, A., Newman, D.: UCI machine learning repository (2007), <http://www.ics.uci.edu/~mllearn/mlrepository.html>
2. Carvalho, R., Freitas, A.: A hybrid decision tree/genetic algorithm method for data mining. *Inf. Sci.* 163(1-3), 13–35 (2004)
3. Chawla, N., Bowyer, K., Hall, L., Kegelmeyer, P.: Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research* 16, 321–357 (2002)
4. Japkowicz, N.: Class imbalances: Are we focusing on the right issue? In: *ICML-KDD'2003 Workshop: Learning from Imbalanced Data Sets* (2003)
5. Klement, W., Wilk, S., Michalowski, W., Matwin, S.: Classifying severely imbalanced data. In: *Proceedings of the 24th Canadian conference on Advances in artificial intelligence*. pp. 258–264. *Canadian AI'11*, Springer-Verlag, Berlin, Heidelberg (2011)
6. Mitchell, T.: *Machine Learning*. McGraw-Hill, New York (1997)
7. Ting, K.M.: The problem of small disjuncts: its remedy in decision trees. In: *In Proceeding of the Tenth Canadian Conference on Artificial Intelligence*. pp. 91–97 (1994)
8. Valiant, L.G.: A theory of the learnable. *Commun. ACM* 27(11), 1134–1142 (1984)
9. Weiss, G.: Mining with rarity: a unifying framework. *SIGKDD Explor. Newsl.* 6(1), 7–19 (2004)
10. WEKA Machine Learning Project: Weka. URL <http://www.cs.waikato.ac.nz/~ml/weka>