The Influence of Class Imbalance on Cost-Sensitive Learning: An Empirical Study

Xu-Ying Liu Zhi-Hua Zhou

National Laboratory for Novel Software Technology
Nanjing University, Nanjing 210093, China
{liuxy, zhouzh}@lamda.nju.edu.cn

Abstract

In real-world applications the number of examples in one class may overwhelm the other class, but the primary interest is usually on the minor class. Cost-sensitive learning has been deeded as a good solution to these class-imbalanced tasks, yet it is not clear how does the class-imbalance affect cost-sensitive classifiers. This paper presents an empirical study using 38 data sets, which discloses that class-imbalance often affects the performance of cost-sensitive classifiers: When the misclassification costs are not seriously unequal, cost-sensitive classifiers generally favor natural class distribution although it might be imbalanced; while when misclassification costs are seriously unequal, a balanced class distribution is more favorable.

1. Introduction

In real-world applications, data sets are often imbalanced, that is, the number of examples in one class may overwhelm the other class. This problem is prevalent in many applications, such as fraud/intrusion detection, medical diagnosis/monitoring, etc. Learning algorithms that do not consider class-imbalance tend to be overwhelmed by the major class and ignore the minor one [3]. However, in class-imbalance learning, usually the primary interest is on the minor class. That is, the cost of misclassifying a minor class example is usually more expensive than that of misclassifying a major one [9, 11].

Various cost-sensitive learning methods [5, 7, 10, 13, 15] have been developed to deal with unequal misclassification costs. Cost-sensitive learning has also been deemed as a good solution to class-imbalance learning [12]. Maloof [8] indicated that, learning from imbalanced data sets and learning with unequal costs can be handled in a similar manner. On one hand, cost-sensitive learning is a good solution to class-imbalance problem [3]; on the other hand, methods designed for class-imbalance problem can also help in cost-sensitive learning [15]. In fact, cost-sensitive learning methods deal with class-imbalance by incurring different costs for different classes. Thus, it is feasible to handle unequal misclassification costs and class-imbalance in a unified framework [14].

However, previous research mainly focus on pure class-imbalance learning or pure cost-sensitive learning, largely ignoring the fact that class-imbalance and unequal misclassification costs usually occur simultaneously. Although it has been observed that class-imbalance has some influence on cost-sensitive classifiers [2, 10, 14, 15], up to now there is no thorough investigation about the influence of class-imbalance on cost-sensitive learning methods.

This paper presents an empirical study using 38 data sets on how class-imbalance would affect cost-sensitive learning methods. The results reveal that class-imbalance does often affect the performance of cost-sensitive classifiers. When costs do not differ seriously, cost-sensitive classifiers generally favor natural class distribution; while when costs differ seriously, cost-sensitive classifiers favor a balanced class distribution instead. This suggests that in the future when we are dealing with a not-seriously imbalanced data set, we can simply apply cost-sensitive learning methods; while when we are facing with a seriously imbalanced data set, before applying cost-sensitive learning methods, we shall try to balance the class distribution.

2. Balancing Class Distribution

We focus on 2-class tasks, where the *positive* class is of the primary interest and with higher misclassification cost C_+, while the other class is the *negative* class with lower cost C_-. There is no cost for correct prediction. Assuming the positive class has N_+ examples and negative class has N_- examples. Since costs can be normalized with the optimal decision unchanged [7], C_- can always be set to 1, and therefore C_+ is always bigger than 1.

A popular approach to cost-sensitive learning is to rescale (or rebalance) the classes such that the influences of different classes on the learning process are in proportion to their costs [14]. A typical process is to assign the training examples of different classes with different weights, where the weights are in proportion to their corresponding misclassification costs, and then pass the weighted examples to a classical learning algorithm such as C4.5 decision tree [7, 10]. Besides weighting the examples, the *Rescaling* approach can also be realized in many other ways, such as sampling the training examples [6–8] or moving the decision thresholds [5,7,15] corresponding to their corresponding misclassification costs.

In *Rescaling*, the optimal rescale ratio (called cost-sensitive rescale ratio) of positive class to negative class is $rc_{+,-}$, as shown in Eq. 1.

$$rc_{+,-} = \frac{C_+}{C_-} \tag{1}$$

In order to deal with class-imbalance using *Rescaling*, different costs are to be incurred for different classes. So, the optimal rescale ratio (called imbalance rescale ratio) of positive class to negative class is $ri_{+,-}$, as shown in Eq. 2, where minor class will have bigger imbalance rescale ratio.

$$ri_{+,-} = \frac{N_-}{N_+} \tag{2}$$

In order to handle unequal misclassification costs and class-imbalance simultaneously, both the cost-sensitive rescale ratio rc and the imbalance rescale ratio ri should be considered. A natural way is to merge them into a single rescale ratio r, as shown in Eq. 3.

$$r_{+,-} = \frac{C_+}{C_-} \times \frac{N_-}{N_+} \tag{3}$$

When minor class has higher misclassification cost and less examples, in order to eliminate the influence of class-imbalance, the rescale ratio becomes larger than that of the case when only misclassification costs are considered. From the view of cost-sensitive learning, the cost of the positive class should be bigger than its given misclassification cost, since the positive class has less examples. That is to say, misclassification cost of a class should be normalized by the number of examples in this class, as shown in Eq. 4.

$$C_+{}' = \frac{C_+}{N_+}, C_-{}' = \frac{C_-}{N_-} \tag{4}$$

Note that Ciraco et al. [4] claimed that in general, changing the class distribution is not equivalent to altering the cost ratio. However, in *Rescaling* they are equivalent, which can be found by comparing Eqs. 3 and 4. Thus, *Rescaling* is a good basis for studying the influence of class-imbalance on cost-sensitive classifiers.

3. Empirical Study

3.1. Settings

In the empirical study, the instance-weighting-based cost-sensitive C4.5 decision tree (denoted by C45CS) [10] is used. In order to handle class-imbalance in cost-sensitive classifiers, the strategy described in Section 2 is employed, which is denoted by B-C45CS.

Thirty-eight UCI data sets [1] are used. Information on these data sets is shown in Table 1. The first part of Table 1 shows eleven imbalanced 2-class data sets; the second part shows twenty-seven 2-class data sets derived from multi-class data sets or relatively balanced 2-class data sets according to one of the following strategies: 1) treating one class of a multi-class data set as the positive class while the union of all other classes as the negative class (e.g. *ann0*), or 2) selecting two classes of a multi-class data set, where one class is regarded as the positive class while the other is regarded as the negative class (e.g. *balance1*), or 3) sampling a small subset from one class of a relatively balanced 2-class data set, which is regarded as the positive class while the other class is regarded as the negative class (e.g. *adult* and *spamebase*). The *imbalance level* (the ratio of major class size to minor class size) in these data sets varies from 1.9 to 66.

The misclassification cost C_- is always set to 1, and C_+ is set to an integer varying from 1 to $max(30, 3 \times imbalance\_level)$. For each $[C_+, C_-]$ combination, five times 10-fold stratified cross validation are performed, and the average results are recorded.

3.2 Results

In order to investigate the influence of class-imbalance on cost-sensitive classifiers, the difference between the classifiers' performance without/with considering class-imbalance is studied.

3.2.1 Performance Tendency

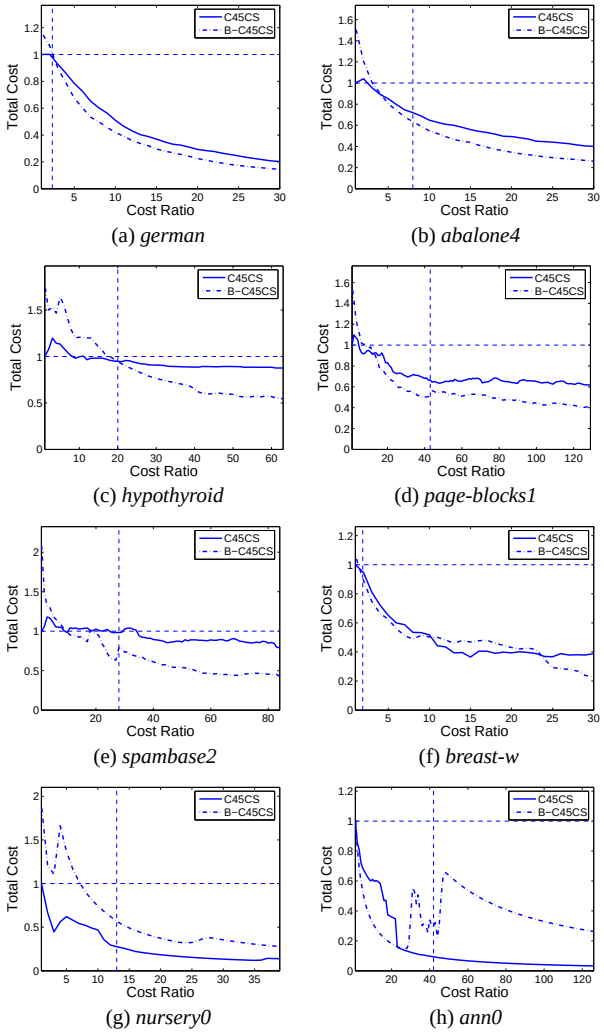
Figure 1 depicts the performance tendency of C45CS and B-C45CS with the increasing cost ratio (ratio of C_+ to C_-) on 8 representative data sets, where the y-axes is the ratio of the classifier's total cost to its corresponding baseline method (standard C4.5). The horizontal dashed line at $y = 1$ indicates the performance of the baseline method. The vertical dashed line points out the imbalance level. Figures 1(a) and 1(b) present typical performance tendency on data sets with slight class-imbalance. The imbalance level of *german* and *abalone4* is 2.3 and 7.5, respectively. Twelve more data sets share such performance tendency, including *breast-cancer*, *credit-g*, *haberman*, *spect*, *abalone2,3,5* *balance1,2*, *adult1*,

Table 1. Experimental Data Sets

Dataset	Size	#Att	#Class	Target	#Maj/#Min
allbp-op	2800	29	2	-	20.1
breast-cancer	286	9	2	-	2.4
breast-w	699	9	2	-	1.9
credit-g	1000	20	2	-	2.3
euthyroid	3163	25	2	-	9.8
german	1000	24	2	-	2.3
haberman	306	3	2	-	2.8
hepatitis	155	19	2	-	3.8
hypothyroid	3163	25	2	-	20.0
sick	3772	29	2	-	15.3
spect	267	22	2	-	3.9
abalone0	4177	8	-	$Age \leq 4$	55.4
abalone1	4177	8	-	$Age \leq 5$	21.1
abalone2	4177	8	-	$Age \leq 6$	8.3
abalone3	4177	8	-	$Age > 12$	5.0
abalone4	4177	8	-	$Age > 13$	7.5
abalone5	4177	8	-	$Age > 14$	10.5
abalone6	4177	8	-	$Age > 15$	15.0
abalone7	4177	8	-	$Age > 16$	20.5
abalone8	4177	8	-	$Age > 17$	29.7
abalone9	4177	8	-	$Age > 18$	43.4
abalone10	4177	8	-	$Age > 19$	66.4
ann0	7200	21	3	Class=1	42.4
ann1	7200	21	3	Class=2	18.6
ann2	7200	21	3	Class=3	12.5
balance0	625	4	3	B	11.8
balance1	337	4	3	L/B	5.9
balance2	337	4	3	B/R	5.9
page-blocks0	5473	10	5	2	15.6
page-blocks1	5028	10	5	1/5	42.7
nursery0	4594	8	4	very/prio	13.0
adult0	12000	14	2	$\leq 50K : 10^4$ $> 50K$:2000	5
adult1	11000	14	2	$\leq 50K : 10^4$ $> 50K$:1000	10
adult2	10500	14	2	$\leq 50K : 10^4$ $> 50K$:500	20
adult3	10200	14	2	$\leq 50K : 10^4$ $> 50K$:200	50
spambase0	3288	57	2	0:2788,1:500	5.6
spambase1	2988	57	2	0:2788,1:200	13.9
spambase2	2888	57	2	0:2788,1:100	27.9

ann1 and *spambase0*. The imbalance level of these data sets are mostly less than 10. Figures 1(c), 1(d) and 1(e) present typical performance tendency on data sets with serious class-imbalance. The imbalance level of *hypothyroid*, *page-blocks1* and *spambase2* is 20.0, 42.7 and 27.9, respectively. Thirteen more data sets share such performance tendency, including *allbp-op*, *abalone0,1,6-10*, *ann2*, *page-blocks0*, *adult2,3* and *spambase2*. The imbalance level of these data sets are all greater than 10.

On data sets described above, C45CS and B-C45CS have



(a) *german* (b) *abalone4*
(c) *hypothyroid* (d) *page-blocks1*
(e) *spambase2* (f) *breast-w*
(g) *nursery0* (h) *ann0*

Figure 1. Performance Tendency

only one cross point. Before the point C45CS is superior to B-C45CS, while after the point C45CS is inferior to B-C45CS. There are some exceptions, as shown in Figure 1(f) where there are more than one cross points. Data sets sharing such performance tendency include *breast-w*, *euthyroid*, *hepatitis* and *sick*. Note that the imbalance level of these data sets is less than 10, and before the first cross point C45CS is superior to B-C45CS. In some cases C45CS always dominates B-C45CS, as shown in Figure 1(g). Data sets sharing such performance tendency include *balance0*, *nrusery0* and *adult0*. There is a very exceptional case (on the data set *ann0*), as shown in Figure 1(h), where there is only one cross point but before the point C45CS is inferior to B-C45CS while after the point C45CS is superior to B-C45CS.

Among the 38 experimental data sets, C45CS and B-C45CS have cross points on 35 data sets. This means that when the minor class has higher misclassification cost,

class-imbalance often affects the cost-sensitive classifier C45CS. But class-imbalance will only take effect when the cost ratio in the concerned task is large enough. When the cost ratio is small, C45CS is generally superior to B-C45CS. Thus, it is not needed to consider class-imbalance in such cases. This is mainly because when the cost ratio is small, considering class-imbalance in cost-sensitive classifiers will make the rescale ratio overly large.

3.2.2 Quantitative Results

To quantitatively study the influence of class-imbalance on cost-sensitive classifiers, min is observed. Figure 2 illustrates this criterion. Since performance tendency shows the performance of a cost-sensitive classifier's against the baseline method, the area under the performance tendency curve (denoted by APT) can be regarded as a performance measure. Smaller APT indicates better performance. In Figure 2, $APT(C45CS) = area(A) + area(B)$, and $APT(B\text{-}C45CS) = area(A) + area(C)$. Thus, when the performance tendency of two methods have cross points, $area(A)$ is always smaller than the APT values of both C45CS and B-C45CS. Otherwise, $area(A)$ equals to the APT value of the dominative classifier. Thus, $area(A)$ is helpful to evaluate the influence of class-imbalance, and its value is defined as min. The value of $APT(C45CS) - min$ indicates how much C45CS favors balanced class distribution: The larger the value, the more the balanced class distribution is favored; while the value of $APT(B\text{-}C45CS) - min$ indicates how much C45CS favors natural class distribution: The larger the value, the more the natural class distribution is favored. Note that all these values should be normalized by the number of cost ratios, since the imbalance level of different data sets are not identical.

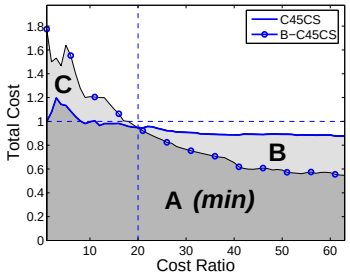


Figure 2. An Illustration of *min*

The total costs of C45CS and B-C45CS are tabulated in Table 2. Here min is also reported. The table entries present the real results of C4.5, and the ratio of the other methods against that of C4.5. Note that, the results for C45CS and B-C45CS are equivalent to the (normalized) APT value. The bolded min values indicates that on the corresponding data set the min value is smaller than the APT value of C45CS, which implies that class-imbalance affects C45CS on that data set.

Table 2. Quantitative Results

Data Set	C4.5	C45CS	B-C45CS	*min*
allbp-op	158.310	.481	.433	**.419**
breast-cancer	99.865	.337	.343	**.325**
breast-w	27.330	.492	.482	**.455**
credit-g	290.445	.361	.344	**.337**
euthyroid	59.805	.670	.719	**.663**
german	255.160	.457	.394	**.386**
haberman	103.835	.345	.346	**.325**
hepatitis	30.235	.500	.494	**.466**
hypothyroid	43.540	.937	.861	**.781**
sick	71.520	.474	.491	**.460**
spect	47.465	.548	.523	**.494**
abalone0	248.340	.538	.221	**.217**
abalone1	274.180	.498	.331	**.317**
abalone2	310.950	.469	.475	**.447**
abalone3	671.020	.562	.502	**.484**
abalone4	573.690	.609	.528	**.503**
abalone5	448.330	.682	.613	**.579**
abalone6	515.400	.663	.515	**.489**
abalone7	572.900	.679	.500	**.473**
abalone8	619.000	.731	.519	**.494**
abalone9	611.080	.799	.529	**.497**
abalone10	616.940	.740	.476	**.455**
ann0	25.310	.156	.382	**.119**
ann1	15.200	.442	.590	**.404**
ann2	22.900	.635	.490	**.488**
balance0	90.650	.675	1.062	.675
balance1	76.390	.491	.407	**.383**
balance2	76.390	.502	.405	**.380**
page-blocks0	90.850	.718	.716	**.645**
page-blocks1	262.440	.706	.572	**.558**
nursery0	169.020	.283	.593	.283
adult0	1695.300	.441	.501	.441
adult1	1050.520	.591	.644	**.556**
adult2	1144.470	.634	.671	**.571**
adult3	1252.430	.735	.746	**.632**
spambase0	172.010	.621	.603	**.589**
spambase1	154.410	.814	.663	**.649**
spambase2	182.110	.939	.689	**.663**
avg.	345.519	.578	.536	.476

Table 2 shows that, the difference between C45CS' APT and min can be as much as 28.5% on data set *abalone10* which suffers from serious class-imbalance (the imbalance level is 66.4). Even on data sets with very slight class-imbalance, min can be smaller than C45CS's APT. For example, on *breast-w*, min is 3.7% smaller than C45CS's APT. The average of min on all data sets is 0.476, which is 10.2% less than the average APT of C45CS.

Generally speaking, class-imbalance does affect the cost-sensitive classifier C45CS. Concretely, C45CS favors natural class distribution in case of small cost ratio, while

favors a balanced class distribution in case of big cost ratio. Moreover, the more serious the class-imbalance, the more significant the influence. Unfortunately, the position of the critical point of C45CS and B-C45CS is hard to be decided since it is task-dependent.

Besides C45CS, we have also empirically investigated the influence of class-imbalance on several other cost-sensitive learning methods, such as instance-weighting-based cost-sensitive Naïve Bayes classifier [13], threshold-moving-based cost-sensitive neural networks [15], and MetaCost [5]. The findings are similar to that have been reported in this paper. Due to the page limit, these results will be presented in a longer version.

4. Concluding Remarks

As far as we know, there is only one work concerned the influence of class-imbalance on cost-sensitive learning [2]. Based on results of three data sets, it was concluded that when misclassification costs are unequal, a training set using the natural class distribution is the best. In this paper, we report on an empirical study involving thirty-eight data sets. Our results disclose that, cost-sensitive classifiers generally favor natural class distribution when costs differ less, while a balanced class distribution is more favorable when costs differ seriously. This only partially agrees with the conclusion of the previous work [2].

It was indicated in [11] that traditional classifiers (i.e. cost-insensitive classifiers) favor a certain class distribution for a certain evaluation metric. Our results disclose that even for the same evaluation metric, a cost-sensitive classifier may favor different class distributions on different data sets. These may suggest that the influence of class-imbalance on traditional classifiers and cost-sensitive classifiers are somewhat different.

In this paper we employ Eq. 3 to merge the cost-sensitive weights as well as class-imbalance weights into a unified framework. It is possible that our results have been biased because of the adoption of this merging scheme. Empirical study with other kinds of merging schemes has been left for future work. Moreover, in this paper we only considers natural class distribution and balanced class distribution. Whether it is better to take distributions other than natural or balanced distributions is an interesting issue to be explored in the future. Furthermore, studying the influence of class-imbalance on multi-class cost-sensitive learning methods is also an interesting future issue.

Acknowledgments: Supported by the National Science Fund for Distinguished Young Scholars of China (60325207), the Jiangsu Science Foundation (BK2004001), and the 973 Fundamental Research Program of China (2002CB312002).

References

[1] C. Blake, E. Keogh, and C. J. Merz. UCI repository of machine learning databases. [http://www.ics.uci.edu/∼mlearn /MLRepository.html], Department of Information and Computer Science, University of California, Irvine, CA, 1998.

[2] P. Chan and S. Stolfo. Toward scalable learning with non-uniform class and cost distributions: A case study in credit card fraud detection. In *Proceeding of the 4th International Conference on Knowledge Discovery and Data Mining*, pages 164–168, New York, NY, 1998.

[3] N. V. Chawla, N. Japkowicz, and A. Kotcz. Editorial to the special issue on learning from imbalanced data sets. *ACM SIGKDD Explorations*, 6(1):1–6, 2004.

[4] M. Ciraco, M. Rogalewski, and G. Weiss. Improving classifier utility by altering the misclassification cost ratio. In *Proceedings of the 1st International Workshop on Utility-Based Data Mining*, pages 46–52, Chicago, IL, 2005.

[5] P. Domingos. MetaCost: A general method for making classifiers cost-sensitive. In *Proceedings of the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 155–164, San Diego, CA, 1999.

[6] C. Drummond and R. C. Holte. C4.5, class imbalance, and cost sensitivity: Why under-sampling beats over-sampling. In *Working Notes of the ICML'03 Workshop on Learning from Imbalanced Data Sets*, Washington, DC, 2003.

[7] C. Elkan. The foundations of cost-senstive learning. In *Proceedings of the 17th International Joint Conference on Artificial Intelligence*, pages 973–978, Seattle, WA, 2001.

[8] M. A. Maloof. Learning when data sets are imbalanced and when costs are unequal and unknown. In *Working Notes of the ICML'03 Workshop on Learning from Imbalanced Data Sets*, Washington, DC, 2003.

[9] D. Margineantu. When does imbalanced data require more than cost-sensitive learning? In *Proceedings of the AAAI'2000 Workshop on Learning from Imbalanced Data Sets*, pages 47–50, Austin, TX, 2000.

[10] K. M. Ting. An instance-weighting method to induce cost-sensitive trees. *IEEE Transactions on Knowledge and Data Engineering*, 14(3):659–665, 2002.

[11] G. Weiss and F. Provost. Learning when training data are costly: The effect of class distribution on tree induction. *Joural of Artificial Intelligence Research*, 19:315–354, 2003.

[12] G. M. Weiss. Mining with rarity - problems and solutions: A unifying framework. *SIGKDD Explorations*, 6(1):7–19, 2004.

[13] B. Zadrozny, J. Langford, and N. Abe. Cost-sensitive learning by cost-proportionate example weighting. In *Proceedings of the 3rd IEEE International Conference on Data Mining*, pages 435–442, Melbourne, FL, 2003.

[14] Z.-H. Zhou and X.-Y. Liu. On multi-class cost-sensitive learning. In *Proceeding of the 21st National Conference on Artificial Intelligence*, pages 567–572, Boston, WA, 2006.

[15] Z.-H. Zhou and X.-Y. Liu. Training cost-sensitive neural networks with methods addressing the class imbalance problem. *IEEE Transactions on Knowledge and Data Engineering*, 18(1):63–77, 2006.