A New Evaluation Measure for Imbalanced Datasets

Cheng G. Weng **Josiah Poon**

School of Information Technologies,
J12, University of Sydney,
Sydney, NSW, Australia 2006,
Email: {cheng, josiah}@it.usyd.edu.au

Abstract

The area of imbalanced datasets is still relatively new, and it is known that the use of overall accuracy is not an appropriate evaluation measure for imbalanced datasets, because of the dominating effect of the majority class. Although, researchers have tried other existing measurements, but there is still no single evaluation measure that work well with imbalanced dataset. In this paper, we introduce a novel measure as a better alternative for evaluating imbalanced dataset. We provide a theoretical background for the new evaluation technique that is designed to cope with cost biases, which changes the previous view about class independent evaluation methods cannot deal with costs, such as ROC curves. We also provide a general guideline for the ideal baseline performance when building classifiers with a known misclassification cost.

Keywords: Evaluation, Imbalanced Datasets, ROC and Cost Sensitive Learning

1 Introduction

Since the workshop at AAAI 2000 (Provost (2000)) , the imbalanced dataset problem has received an increasing attention for the past few years. This area of research focuses on datasets with skewed class distribution and that the minority class is the class of interest. Imbalanced datasets can occur in many domains, such as medical, information technology, biology, and finance. With imbalanced datasets, the conventional way of maximizing overall performance will often fail to learn anything useful about the minority class, because of the dominating effect of the majority class. Consider a problem where 99% of the data belongs to one class, and only 1% is rare class examples. A learner can probably achieve 99% accuracy with ease, but still fail to correctly classify any rare examples. Conventional approaches can produce misleading results on imbalanced dataset, so it is important to know that one needs to take a more localized approach at all levels when dealing with an imbalanced dataset.

Evaluation is the key to making advances in data mining, and it is especially important when the area is still at the early stage of its development. Imbalanced dataset community has criticized the use of non-class independent evaluation measures, such as the overall accuracy, for reporting experimental results on imbalanced datasets. The non-class independent evaluation fails because the results only reflect the learning performance of the majority class, and the more skewed the class distribution is the worse the effect will be. Therefore, when we evaluate the performance on imbalanced datasets, we want to focus on individual classes.

		Predicted class	
		+ve	-ve
Actual Class	+ve	True Positive(TP)	False Negative(FN)
	-ve	False Positive(FP)	True Negative(TN)

Table 1: Confusion matrix for a 2-class problem

There are many evaluation measures in data mining, some of the most relevant ones to imbalanced datasets are: precision, recall, F-measure, Receiver Operating Characteristic (ROC) curve, cost curve and precision-recall curve. The commonality shared between them is that they are all class independent measurements. In particular, ROC curve is well received in the imbalanced dataset community and it is becoming the standard evaluation method in the area. Provost et al. (1998) have argued that reporting accuracy can be misleading, but ROC curve can help explore different tradeoff among different classifiers over a range of operating conditions. However, using ROC curve is hard to compare different classifiers for different misclassification cost and class distributions

In this paper, we demonstrate a generalize form, base on different cost bias, of computing the area under ROC curve (AUC), which we will refer to as weighted-AUC. We will describe the related evaluation measurements for imbalanced dataset in related works and present the details of the weighted-AUC in section 3. We show that weighted-AUC is a better alternative when evaluating imbalanced datasets. In section 4, we will define what the ideal baseline performance should be when the misclassification cost is given. Next, we present some experimental results comparing the normal AUC and weighted AUC. Finally, we will discuss some issues related to weighted-AUC in the discussion section.

In the rest of the paper, minority or positive may be used to refer to the rare class and majority or negative for reference to the common class. The examples in this paper will be restricted to two-class problems.

2 Related Works

In imbalanced datasets, not only is the class distribution is skewed, the misclassification cost is often uneven too. The minority class examples are often more important than the majority class examples. The cost of misclassify a minority class example is

2(a)		Predicted class	
		+ve	-ve
Actual	+ve	0	1
Class	-ve	1	0

(a) Equal cost case

2(b)		Predicted class	
		+ve	-ve
Actual	+ve	0	5
Class	-ve	1	0

(b) Uneven cost case

Table 2: Cost Matrix Examples

far greater than misclassify a majority class example, for example, fraud detection or cancer diagnosis. We will briefly discuss some relevant evaluation measurements, starting with confusion matrix, which is closely related to many evaluation techniques and can be found in most data mining textbooks. Table 1 shows a confusion matrix for outcome of a two class problem. As an example, using this table, we can define the overall accuracy as $\frac{TP+TN}{TP+FP+TN+FN}$. Confusion matrix is useful when accessing the performance without taking cost in to consideration. It is used as a basis for various measures, such as precision and recall.

Precision, recall and F-measure In information retrieval, if there is Z number of relevant documents in the database, and a system can returns X number of documents in which only Y number of them are relevant. We can compare different system performance with precision and recall, which are defined as eq.1 and 2. They can also be defined using confusion matrix terms.

$$Precision = \frac{X}{Y} \texttt{ or } \frac{TP}{TP+FP} \tag{1}$$

$$Recall = \frac{Y}{Z} \texttt{ or } \frac{TP}{TP+FN} \tag{2}$$

F-measure is a common evaluation metrics that combines precision and recall into a single value, usually with equal weighting on both measures.

$$F-measure = \frac{2 \times Recall \times Precision}{Recall + Precision} \tag{3}$$

Cost matrix Sometimes, the costs are known for the problem at hand, i.e. the misclassification cost of a positive or negative example. In this case, we can use the known cost to penalize the resulting confusion matrix to arrive at a meaningful performance assessment. A cost matrix looks the same as a confusion matrix, except it show the cost of misclassification. Table 2(a) shown an general equal cost case, where 2(b) shows an uneven cost matrix of 1 to 5, which is quite normal for imbalanced datasets. When evaluating, the cost is multiplied on top of a confusion matrix to reflect the true performance.

The disadvantage of using confusion matrix-based evaluation is that they are only looking at the performance on a "spot", which means we cannot tell how different class distribution or different cost will affect the performance. So researchers may prefer to visually see the performance over a range of situations using one of the graphical evaluation tools, such as a ROC curve.

ROC and AUC Receiver Operating Characteristic (ROC) curve is a common evaluation technique, originated from radio signal analysis (Green and Swets (1966)) and introduced to machine learning by Spackman (1989) and made popular in data mining by Provost and Fawcett (1997). ROC curve presents the tradeoff between the true positive rate and the false positive rate; as a learner captures more positive example, it will generally misclassify more negative examples as positive examples. If we have a two class problem, a ROC curve can be plotted by varying the probability threshold, from 0 to 1, for predicting positive examples. The true positive rate is the same as recall and the false positive rate equals to $\frac{FP}{FP+TN}$.

A ROC curve is consider to be good if it is closer to the top left corner, and the straight line connecting (0,0) and (1,1) represents a random classifier with even odds. The advantage of use ROC is that one can visually see for what region a model is more superior compare to another.

The area under the ROC curve (AUC) is often used to summaries a learner's performance into a single quantity, which represents the performance of a learner in general across different prediction cost, and the larger the AUC the better. However, ROC-based measures still lack support for tasks involving uneven cost matrix. In addition, it is interesting to know that all the confusion matrix-based measures can be seen as a single point on the ROC curve.

Cost curve Cost curve was introduced by Drummond and Holte (2000), and they have also provided a detailed comparison between ROC curve and cost curve in Drummond and Holte (2004). Basically, cost curve looks at how classifiers perform across a range of different misclassification cost. It can be seen as different slope line tangent to the ROC curve, therefore every ROC curve has a corresponding cost curve. This view of a slope line bears similarity to the discussion about baseline performance in section 4.

Precision and recall (PR) curve Information retrieval experts use PR curve in similar fashion as ROC curve – except that because the axes are different and the ideal classifier is toward the upper right. An example is shown in figure 7. Davis and Goadrich (2006) provided a comparison between PR curve and ROC curve.

All the graphical evaluation tools provide different perspectives to access a learner's performance, and the advantage is to have a better understanding of the learner's behavior under a range of circumstances.

However, the measurements described in this section are designed to work well for different purpose; for example, ROC is well suited for ranking problems with a goal to achieve best resources utilization, whereas cost curve is suitable to locate best method for certain operating cost constraint. Therefore, we realized there is a need to design an appropriate evaluation method for dealing with imbalanced dataset, hence we propose a general approach that can take cost bias into account at evaluation

3 Weighted AUC

An imbalanced dataset often has different cost matrix than a balanced dataset, this is generally due to the nature of the imbalanced dataset, e.g. detection oil spill on satellite images, scanning for scam websites, or detecting cancer. The misclassification of a rare

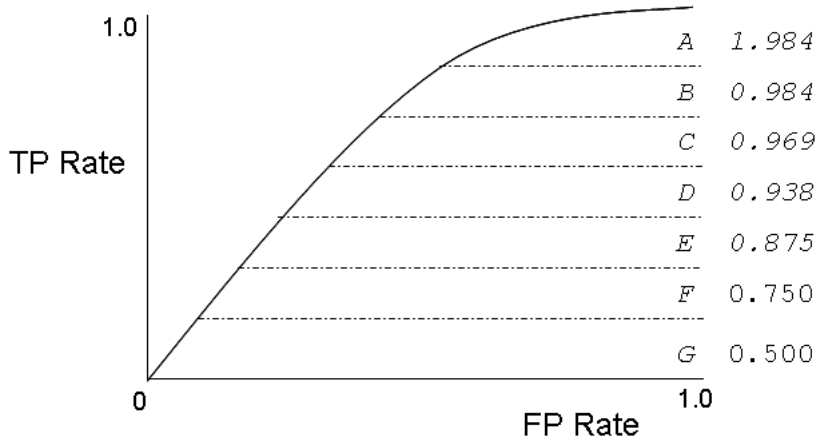

Figure 2: Weighted-AUC example: new weight vector after 50% weight transfer

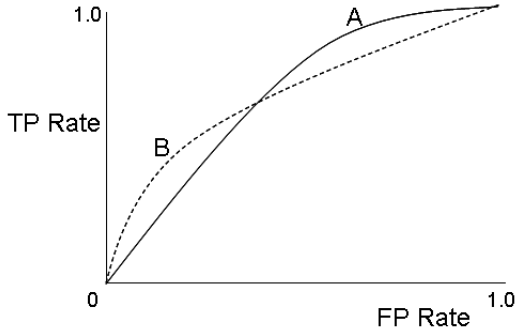

Figure 1: Compare classifiers with ROC curve when conventional AUC is the same.

occurring case in these situations can be very costly, much more expensive than a false alarm. Hence, there is a tendency for the imbalanced datasets to bias its cost towards positive examples. Therefore, one may require the learners to achieve a certain performance level by presetting the cost matrix. As an example, say detecting cancers from x-ray scans, the learners built must able to achieve a standard recall rate in order to be declare useful, otherwise it will not be used no matter how high the precision is. So if the target recall is 90%, then comparison of performances should only consider the area above the 90% TP rate on the ROC curve. In section 4, we provide more detail discussion about the relationship between cost matrix, class distribution and baseline performance.

The rationale for the weighted-AUC measurement is based on the notion that when one is dealing with imbalanced datasets, a learner that performs well in the higher TP rate region is preferred over ones that does not. In another word, a false negative is worse than a false positive. So, in the ideal case, the learner should be able to catch every positive example, i.e. 100% recall/TP rate. With this ideal in mind, if one wants a 100% recall learner, then the best choice will be the one that has the lowest FP rate at the 100% TP rate line.

Generally, where the cost is all equal and two learners have the same AUC, as in figure 1, one can not say classifier A is better than classifier B. However, in an imbalanced dataset situation, one can visually see that classifier A is more appropriate, because at higher TP rate region classifier A has smaller FP rate than classifier B. So, classifier A is preferred over classifier B even though they have the same AUC value. Therefore, the problem with conventional AUC is that it does not consider cost bias, because we sum up the areas with equal weights of 1, which is a fair assumption when we have equal cost.

We propose a skewed weight distribution method that allows one to compute AUC with a cost bias. We refer to this approach as weighted-AUC. When the cost is uneven and biased towards the rare class, instead of summing up areas with equal weights, we want to give more weights to the areas near the top of the graph. So we create a skew weight vector by distributing more weights towards the top of the ROC curve, while keeping the total weights unchanged. The idea is to pass certain percentage of weights from the bottom areas towards the upper areas of the ROC curve. In figure 2, we present the resulting weight vector after a 50% weight shift is performed. Originally all weights were 1, but now 50% of area G is passed to area F, and 50% of the new area F's weights is again passed to area E and so on. The weight shifting process stops at the top, area A, which has the most weight. We can define the new weights more formally with eq 4. If we have N number of areas to sum:

$$W(x) = \begin{cases} \alpha, & x = 0 \\ W(x-1) \times \alpha + (1-\alpha), & 0 < x < N \\ \frac{W(x-1)\times\alpha+(1-\alpha)}{1-\alpha}, & x = N \end{cases} \tag{4}$$

Where α is the percentage of weight to transfer to the next area towards the top. α ranges from 0, no weight transfer, to 1, a total weight transfer. When α is 0, the resulting weighted-AUC is equal to the conventional AUC; when α is 1, only the area at the top is considered. $W(0)$is the weight for the bottom area. The new weight of an area is defined as a recursive formula using the weight of the previous area. This new weight vector is used to compute a new AUC value by adding up the areas times their corresponding new weight.

$$\texttt{weighted} - \texttt{AUC} = \sum_{i=0}^{N} area(i) \times W(i)$$

If the cost is known, we can use the cost ratio between the positive and negative classes to set the weight transfer rate, α.

$$\alpha = 1 - cost\,ratio$$

The maximum and minimum of weighted-AUC is the same as original AUC, 1 and 0. The advantage and disadvantages of weighted-AUC is similar to the

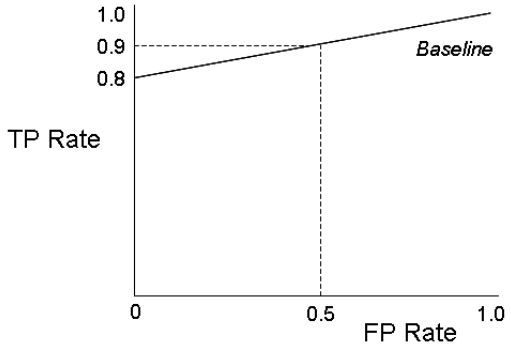

Figure 3: The ideal baseline performance

conventional AUC, except weighted-AUC is enhanced with the ability to adjust to different cost bias. This challenges the traditional view of ROC and AUC, where they were considered to be cost inconsiderate evaluation measures (Drummond and Holte (2004)).

4 Ideal Baseline Performance

If the misclassification cost is known then one can define the baseline performance for the learners. Suppose the cost is same as in table 2(b), and if we assume the future class distribution is balanced, then the baseline performance is shown in figure.d. For example, if we need to classify 100 positive and 100 negative examples, with the cost of 1 to 5, when we classify everything as positive, the cost is 100. In order to match the same cost of 100, we can only afford to miss out 20 positive cases ($1/5 = 20\%$), while we correctly classify all negative examples, meaning zero FP rate. Following this reasoning, we can set the baseline performance by drawing a straight line connecting (80% TP rate, 0% FP rate) and (100% TP rate, 100% FP rate). A classifier is worth considering if it can achieve a performance above this line.

We have assumed equal class distribution in the above case for simplicity sake, but when we have an imbalanced class distribution, the baseline performance may need to be adjusted because the future distribution could also be skewed. If we use the training data class distribution as an estimate for the future class distribution, then the adjustment to the baseline performance is done by dividing the cost ratio over class distribution ratio. For example, if we have an imbalanced dataset with 100 positive and 400 negative examples, then with the cost of 1 to 5, the adjusted baseline should equal to$(1/5)/(100/400) = 0.8$. This means if we correctly classify every negative examples, then we can afford to miss 80 positive examples. The adjusted baseline performance will then go through (20%,0%) and (100%,100%).

It is important to know that when setting the misclassification cost, one should take the estimated class distribution into consideration, because they are closely related. If the class distribution ratio were to be less than the cost ratio, then the adjusted baseline performance will actually bias towards the negative class. Therefore, when one tries to set the cost ratio for an imbalanced dataset, where the rare class example is more important, one should generally consider have the smaller cost ratio than the estimated class distribution ratio.

It is interesting to note that the concept of cost curve is also about creating slopes, but instead of drawing slopes based on cost, the cost curve draws the slope lines tangent to a ROC curve to reflect the performance of a particular operating point.

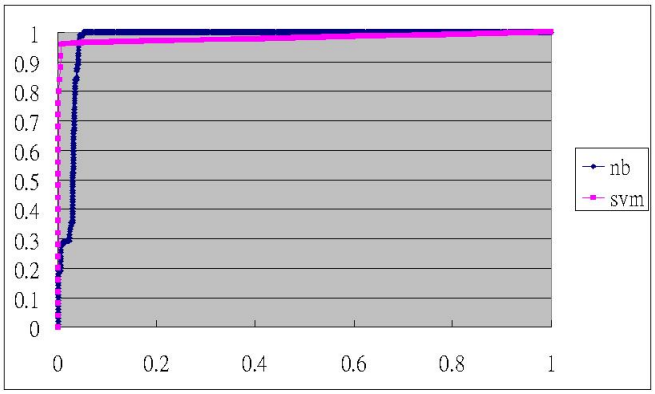

Figure 4: ROC curve for *nb* and *svm* learning on 'anneal-2' dataset

We can see that *nb* touches the 100% TP rate much early than *svm* does, even though *svm* performs better at lower FP rate, and when working with imbalanced dataset, it is more desirable to have a learner that touches 100% TP rate at that lowest FP rate. So *nb* should be considered as having a better ROC curve for imbalanced datasets with cost bias towards the minority class.

5 Experiments

After introducing the background theoretical motivations for a weighted AUC, it would not be complete without looking at some real numbers from experiments. So, we have conducted an experiment to compare the normal AUC value and the weighted-AUC values. Four different learners were used in our experiment: Naïve Bayes(*nb*), Decision tree(*j48*), Support vector machines(*svm*), and k-nearest neighbour(*3nn*, we set the k to 3). The datasets are from UCI data repository (A. Asuncion, 2007) and we took multi-class problems and turned them into binary classification problems by treating one of the class as the positive class and use the rest of the classes as negative class. The process created 98 datasets.

When we compare the AUC value along side with weighted-AUC, it is clear that the more superior learner under AUC evaluation is not necessary better under weighted-AUC assessment. Table.3 shows both AUC and weighted-AUC values for a sample of 20 datasets out of 98 datasets across 4 different learners. We only shown 20 because it is sufficient to point out the difference between normal and weighted AUC. We have used 0.1 for the α, meaning we will transfer 10% of the weights.

The datasets in bold are where learners' performance ranking differs when evaluated under different metrics. To demonstrate, we look at the first occurrence of conflicting performance ranking, which happened with '*anneal-2*' dataset. For this dataset, when ranking the learning performance by the original AUC values, the learners' rank were *j48*, *nb* and *svm*, and follow by *3nn*. However, when we look at weighted-AUC for the same dataset and learners, the ranking changed to *nb*, *j48*, *svm*, and *3nn*. So *nb* becomes the best learner out of the four. If we take a closer look at the ROC curves, as shown in figure.4, where we try to compare *nb* and *svm*, since they has the same normal AUC value, but different weighted-AUV value. We can see that *nb* does have a much better ROC curve for imbalanced dataset than *svm*, because it touches 100% TP rate much earlier than *svm*, which is not shown by the normal AUC.

Dataset	normal AUC				weighted-AUC			
	nb	j48	smo	3nn	nb	j48	smo	3nn
anneal-1	0.99	0.64	0.94	0.82	0.9	0.61	0.85	0.75
anneal-2	0.97	1	0.97	0.89	0.97	0.9	0.88	0.86
anneal-5	1	0.99	1	1	0.98	0.89	0.9	0.9
audiology-age	0.98	0.98	0.95	0.92	0.96	0.89	0.86	0.89
audiology-age_and_noise	0.99	0.96	0.92	0.63	0.94	0.87	0.84	0.6
audiology-cochlear_unknow	0.91	0.92	0.89	0.81	0.89	0.88	0.82	0.78
audiology-poss-noise	0.98	0.78	0.9	0.87	0.94	0.73	0.83	0.81
autos-2	0.65	0.81	0.63	0.81	0.64	0.77	0.63	0.78
balance-scale-B	0.33	0.5	0.5	0.35	0.33	0.5	0.5	0.35
balance-scale-L	0.99	0.84	0.92	0.98	0.99	0.81	0.83	0.97
balance-scale-R	0.99	0.84	0.92	0.98	0.99	0.81	0.83	0.97
breast-cancer	0.7	0.61	0.58	0.64	0.69	0.6	0.58	0.63
cleveland-heart-50_1	0.91	0.77	0.83	0.85	0.9	0.73	0.77	0.81
credit-rating	0.9	0.89	0.86	0.75	0.89	0.88	0.78	0.72
german_credit	0.79	0.65	0.67	0.61	0.78	0.63	0.65	0.6
Glass-buildwindfloat	0.76	0.81	0.57	0.87	0.75	0.78	0.58	0.82
Glass-buildwindnonfloat	0.7	0.76	0.5	0.84	0.69	0.73	0.5	0.8
heart-statlog	0.9	0.79	0.83	0.84	0.89	0.76	0.77	0.79
hepatitis	0.86	0.67	0.77	0.76	0.83	0.64	0.73	0.74
horse-colic.ORIG	0.79	0.5	0.71	0.61	0.78	0.5	0.68	0.59

Table 3: Compare normal AUC with weighted-AUC for UCI datasets.
We show a sample of the 98 datasets we have, and we highlight the datasets in bold to show where learners' performance ranking differs under different evaluation methods, i.e. normal and weighted AUCs.

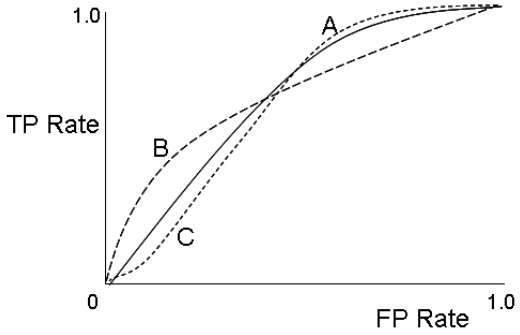

Figure 5: Same weighted-AUC example

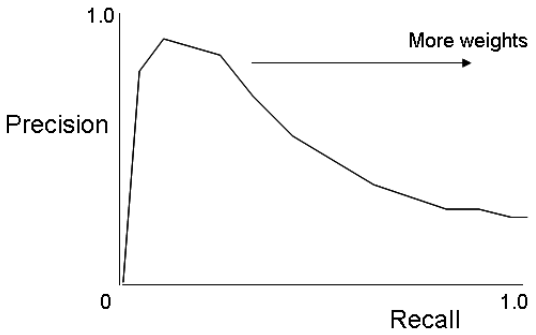

Figure 7: Precision-Recall curve

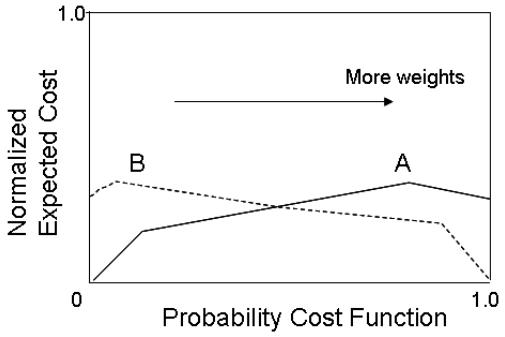

Figure 6: Cost curve

6 Discussion

There is an interesting similarity between cost curve and weighted-AUC, in which they are both designed to work with cost. However, cost curve is for visualizing performance over a range of costs, whereas the purpose of weighted-AUC is to use a biased weight vector to give an appropriate summarized quantity without sacrifice the advantage of graphical evaluation methods. Weighted-AUC is a compromise between a single value and a graphical view. The same weight vector idea can be applied on different graphical evaluation methods, such as cost curve and precision-recall curve. For a cost curve, as shown in figure 6, the weight shifts towards the right side of the graph if the cost bias is towards the positive class. For the same cost preference, the weight also shifts towards the right side for a precision-recall curve, as shown in figure 7.

Some may argue, weighted-AUC will still have the same problem as the conventional AUC when the learners have equal weighted-AUC value, but have different ROC curves. This phenomena is illustrated in figure 5, where classifier A and classifier C have the same weighted-AUC. However, the purpose of weighted-AUC is to achieve cost bias evaluation, which means we can separate classifier A and B. It is alright to have two different learners both perform equally well under the same cost constraint.

When using weighted-AUC, one needs to set the a for the percentage of weight to transfer. This adjustable parameter allows the flexibility of adapting to different cost bias when the cost is known. In the case of unknown cost, one can always use the class distribution ratio as an estimate for the cost ratio. Weighted-AUC can be considered as a generalized form of AUC in terms of different cost biases.

Based on the concept of weighted-AUC, it can provide new insights for previous researches. For example, in a study by Weiss and Provost (2001), where they experimentally show that when conventional AUC is used as performance measure, then a balanced

class distribution should be used for training. This finding conforms with the concept of weighted-AUC, which equates that the conventional AUC as having zero weight shift; this translate to assuming zero cost bias or a balanced class distribution. Therefore, because the conventional AUC was used in the evaluation for their experiment, so only balanced class distribution will give the best results under the conventional AUC. It is like having a normally distributed graph and you want to find another bell-shaped graph that will yield the maximum overlap between the two, which will inevitably land you on another normally distributed graph with the same

7 Conclusion

We have introduced a new evaluation method for imbalanced datasets, called weighted-AUC. One can think of it as an enhanced version of AUC, a measure that is already gaining popularity in the imbalanced dataset community. We have shown why weighted-AUC is a better alternative under cost biased situations. A discussion for setting misclassification cost, baseline performance for imbalanced datasets is provided as a general guideline when dealing with imbalanced datasets. In the end of discussion section, we have also presented an example of how weighted-AUC can give new sights for researches in the data mining area. In the future, when one is dealing with imbalanced datasets, we recommend the use of weighted-AUC in place of conventional AUC in order to give a better cost-biased comparison.

References

A. Asuncion, D. N. (2007), 'UCI machine learning repository'.
URL: *http://www.ics.uci.edu/∼mlearn/MLRepository.html*

Davis, J. and Goadrich, M. (2006), The relationship between precision-recall and roc curves, *in* 'In ICML06: Proceedings of the 23rd international conference on Machine learning', pp. 233–240.

Drummond, C. and Holte, R. C. (2000), Exploiting the Cost (In)sensitivity of Decision Tree Splitting Criteria, *in* 'Proceedings of the Seventeenth International Conference on Machine Learning', pp. 239–249.

Drummond, C. and Holte, R. C. (2004), What roc curves can't do (and cost curves can), *in* 'ROCAI', pp. 19–26.

Green, D. and Swets, J. (1966), *Signal detection theory and psychophysics*, John Wiley and Sons Inc.

Provost, F. (2000), Machine Learning from Imbalanced Data Sets 101, *in* 'AAAI Workshop on Learning from Imbalanced Data Sets', AAAI Press.

Provost, F., Fawcett, T. and Kohavi, R. (1998), The case against accuracy estimation for comparing induction algorithms, *in* 'Proceedings of the Fifteenth International Conference on Machine Learning', pp. 43–48.

Provost, F. J. and Fawcett, T. (1997), Analysis and visualization of classifier performance: comparison under imprecise class and cost distributions, *in* 'Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining', pp. 43–48.

Spackman, K. A. (1989), Signal detection theory: Valuable tools for evaluating inductive learning, *in* 'Proceedings of the Sixth International Workshop on Machine Learning', pp. 160–163.

Weiss, G. M. and Provost, F. (2001), The Effect of Class Distribution on Classifier Learning: An Empirical Study, Technical report, Department of Computer Science, Rutgers University. Technical Report ML-TR-44.