Class Probabilities and the ROC Curve in Changing Environments

William Klement KLEMENT@SITE.UOTTAWA.CA
Nathalie Japkowicz NAT@SITE.UOTTAWA.CA
Stan Matwin STAN@SITE.UOTTAWA.CA
University of Ottawa, Ottawa, Ontario, K1N 6N5 Canada.

Keywords: Classification, Probability Estimates, Concept Drift, ROC Analysis

Abstract

A probabilistic classifier assigns probability scores to data examples. The ROC curve depicts the ranking performance of the classifier by imposing threshold values on these probabilities and by plotting the true positive rates against the false positive rates. Once classification decisions are made, the ROC eliminates these probabilities from the performance analysis. In this paper, we argue that discarding these probabilities can result in a loss of information. We show that the lost information is related to changes in the domain. We propose an evaluation method, the sensibility analysis, to remedy this situation. Using synthetic data, we illustrate that visualizing the sensibility values for all classification thresholds depicts changes in the underlying distribution, otherwise known as concept drift.

1. Introduction

For a probabilistic classifier in a binary classification problem, the ROC curve is generated by plotting the true positive rates against false positive rates for all classification thresholds between 0 and 1. The true positive and the false positive rates are obtained by imposing the classification thresholds on class membership probabilities, assigned to data examples by the classifier, to produce a confusion matrix generating points that form the ROC curve (Provost & Fawcett, 2001). Once classification decisions are made, the probability scores are excluded form the performance analysis. In a last year's workshop (Klement & Flach, 2008), we proposed a method to incorporate these probability scores into the ROC space based on the intuition that the probability scores provide additional insights beneficial to the assessment of the classifier's performance. We argued that the ROC analysis fails to distinguish between examples whose scores differ in magnitude. For instance, if a probabilistic classifier assigns probabilities 0.9 and 0.51 to two positive examples respectively, a classification threshold of 0.5 results in both examples being classified correctly, however, the margin in their probabilities remains ignored.

In this paper, we present a scenario of when class membership probabilities present additional information related to the data distribution. Further, we propose the concept of classification sensibility which, we argue, is capable of detecting that what the ROC analysis fails to detect. In particular, we use a synthetic medical data generator to introduce changes in the underlying distribution. We train a Naive Bayes classifier data drawn from the original synthetic domain and test it on data drawn from the modified domain with various types of concept drift (Widmer & Kubat, 1996). Our illustration shows that, in some situations, the ROC curve of the training data remains indistinguishable for that of the testing data despite a change in the domain. Our sensibility curve, on the other hand, is capable of visualizing such changes.

Recently, machine learning assumptions have been criticized (Hand, 2006) of being ignorant to changes in the underlying data distributions. The criticism is based on the assumption which, most machine learning methods make, that training and testing data are drawn form the same static distribution. Hand argues that training and testing data may be drawn from different data distributions more often than not. This problem of learning in changing environments has been studied by machine learning researchers for

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

over a decade (Alaíz-Rodríguez & Japkowicz, 2008; Klinkenberg, 2004; Narasimhamurthy & Kuncheva, 2007; Widmer & Kubat, 1996). In this work, the focus is on the ROC analysis in the context of changing environments. We present simulation of changing medical environments (Alaíz-Rodríguez & Japkowicz, 2008) in which, we show that the ROC my be unable to detect changes in the domain. To remedy this situation, we propose the utilization of those probability scores, discussed earlier, which are excluded from the ROC analysis altogether.

2. Classification Sensibility

The principle of sensibility is to measure classification performance on two subsets of the data. These two subsets are the sensible and the non-sensible examples respectively. As shown in table 1, this division is based on calculating a midpoint t_s which we describe first. Later, we illustrate how we compute the sensibility, the capability and the struggle ratio metrics.

Let X be a set of n examples where the i^{th} example is a vector x_i of values for attributes $a_1, a_2, \ldots, a_m$. Each x_i is assigned a label $c_i \in C = \{+, -\}$ in a binary classification problem for simplicity. Let p_i^+ and p_j^- be the probabilities assigned to the positive example i and to the negative example j respectively. Let n^+ and n^- be the number of positives and negatives in X also respectively. For ease of notations, let $P^+ = \sum p_i^+$ and $P^- = \sum p_i^-$, thus, $P = P^+ + P^-$, be the sum of all probabilities assigned to X. Then, we can calculate the mean scores of positives $m^+ = \frac{P^+}{n^+}$ and the mean score of negatives $m^- = \frac{P^-}{n^-}$ respectively. The class distribution is $d = \frac{n^+}{n^-}$ and in the case of calibrated p_i scores, $m^+ + \frac{m^-}{d} = 1$. This is obvious in the extreme case where $p_i \in \{0, 1\}$ and $d = 1$ where $n^+ \times m^+ + n^- \times m^- = n^+$. Generally, the p_i scores are not calibrated, and the class distribution $d \neq 1$, then we have $m^+ + \frac{m^-}{d} = \frac{P^+}{n^+} + \frac{1}{d} \cdot \frac{P^-}{n^-}$. Therefore, $m^+ + \frac{m^-}{d} = \frac{P^+}{n^+} + \frac{n^-}{n^+} \cdot \frac{P^-}{n^-} = \frac{P}{n^+}$. The midpoint t_s for X is estimated by:

$$t_s = \frac{1}{2}(m^+ + \frac{m^-}{d}) = \frac{1}{2} \times \frac{P}{n^+} \quad (1)$$

We illustrate this concept in table 1. We propose that *sensible* examples are either, positives whose probabilities p_i are above the midpoint t_s, or, negatives that are assigned probabilities p_i below the midpoint t_s. With a midpoint $t_s = 0.56$ and a classification threshold of 0.35, this division appears in the right column of table 1. The positive examples $(1, 2, 4, 5)$ are assigned p_i probabilities above the midpoint $t_s = 0.56$. The negative examples $(6, 8, 9, 10)$ are assigned p_i probabilities below the midpoint $t_s = 0.56$. The union of these examples forms the set of sensible examples. On the other hand, the positive example (7), its p_i is below the midpoint $t_s = 0.56$, as well as, the negative example (3), its p_i is above the midpoint $t_s = 0.56$, form the set of non-sensible (or difficult examples). This is due to the disagreement between their probability assignments, relative to the midpoint t_s, and their labels. Table 1 shows that the classifier assigns sensible probabilities to examples 1,2,4,5,6,8,9, and 10. Examples 3 and 7, on the other hand, are assigned non-sensible probabilities. It is important to mention that the probabilities are assumed to reflect positive class membership expressed by a probabilistic classifier.

Table 1. Sensible and non-sensible examples for classification threshold =0.35 and a midpoint = 0.56.

i	Label	p_i	Prediction	Sensible?
1	+	1.0	+	yes
2	+	0.9	+	yes
3	-	0.8	+	no
4	+	0.7	+	yes
5	+	0.6	+	yes
6	-	0.5	+	yes
7	+	0.4	+	no
8	-	0.3	-	yes
9	-	0.2	-	yes
10	-	0.0	-	yes

The principle of sensibility analysis is based on distinguishing between classification errors made on these two subsets of examples. From an evaluation perspective, we would expect a sensible classifier to make few errors on the set of sensible examples because their probabilities are sensible. In addition, a classifier that makes few errors on the set of non-sensible examples is a capable classifier, i.e. capable of addressing the non-sensible probability assignment. The over all idea is to distinguish between errors made due to wrong classifications from those due to non-sensible probability assignment.

We now describe how to compute our performance metrics. First, we measure the extent to which the probability estimation struggles with the given data. The fraction of non-sensible to sensible examples shows just that, we call it, the *struggle ratio*. In table 1, the struggle ratio is $\frac{2}{8} = 0.25$. Second, the classification accuracy on the set of sensible examples measures classification *sensibility* and, in our example, is 7 out of 8 sensible examples (6 is an incorrectly classified sensible example) for a classification threshold of 0.35. There-

Table 2. An illustration with synthetic flu symptoms.

Data Set	AUC	Midpoint t_s	Struggle Ratio
Original Distribution			
NGP-Train	0.815	0.512	0.488
NGP-Test	0.800	0.518	0.486
Modified Distribution			
DP-Test	0.770	0.338	0.691
FC-Test	0.803	0.681	0.351

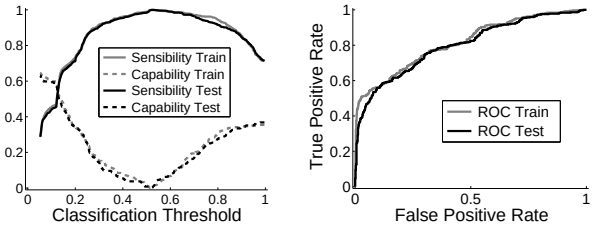

Figure 1. Left: Sensibility and Capability curves. Right: the ROC Curve for training and testing NGP domain

fore, the sensibility for table 1 is $\frac{7}{8} \times 100 = 87.5\%$. Similarly, the classification accuracy on the set of non-sensible examples depicts the capability of the classifier in classifying these difficult examples. In table 1 and using a classification threshold of 0.35, the capability is 1 out of 2 non-sensible examples (3 is an incorrectly classified non-sensible example) resulting in capability of $\frac{1}{2} \times 100 = 50\%$. Finally, for a given data, we can only compute one midpoint t_s and, thus, one struggle ratio. This is so because there is is only one probability score assigned to each example. The sensibility and the capability values, however, vary depending on the classification threshold. For these, we generate their values for all classification thresholds (between 0 and 1) in a manner similar to generating ROC curves (Provost & Fawcett, 2001).

3. Illustrations with synthetic data

To demonstrate information added by considering the probabilities in the evaluation process, we use a synthetic medical data generator to introduce changes to the domain for testing. In each case of testing, we plot the corresponding curves for the sensibility, the capability, and the ROC . The synthetic medical domain (Alaíz-Rodríguez & Japkowicz, 2008) models the prognosis of patients, infected with the flu symptoms, as *NormalRemission* or *Complication*. The original domain corresponds to the *Negative Growth Population* (NGP) with several attribute dependencies. We train a Naive Bayes classifier on data generated from the original domain NGP, then, test it on three data sets. We test on the original NGP domain, i.e. on the training data and on test data obtained from the original unmodified NGP domain. Then, we test the Naive Bays classifier, trained as before, on test data obtained from the Developing Population domain, the DP domain. This domain is a modified NGP domain representing the population in a developing country region where the medical classifier is deployed (Alaíz-Rodríguez & Japkowicz, 2008). This domain contains a population drift. The third and final test evaluates the Naive Bayes classifier, again trained as before, on the FC data. This represent a modified version of the original NGP domain by introducing a class definition change with fewer complications. We generate 1000 examples for each data set. Further details can be found in (Alaíz-Rodríguez & Japkowicz, 2008).

Table 2 illustrates that, when test data is drawn from the original domain NGP, for both training and testing, the AUC value is over 80% and the struggle ratio is under 50%. The AUC and the struggle ratio values remain unchanged for testing on the training set and on the testing set. This is also depicted by the corresponding sensibility, capability and ROC curves in figure 1. All three types of curves remain unchanged between training and testing. In this case, both training and testing data are obtained form the original NGP domain. This leads us to the conclusion that the ROC and our sensibility curves depict the same results when data is drawn from the same static distribution. However, if we consider the Developing Population DP data, the corresponding AUC value in table 2 shows a drop to 77%. The Struggle ratio increases to 69% indicating a higher struggle in the probability estimation. Inspecting the plots in figure 2 reveals that all curves, corresponding to testing on DP data, visually differ from those obtained by testing on the training data. This corresponds to the fact that the DP test data is generated with a population drift. The observed drop in performance and the increased struggle can be attributed to the change in the testing domain.

However, results obtained by testing on the FC data shows that, while the AUC value does not change, the struggle ratio deceases to 35%. This implies that the probability estimation struggles less with fewer complications in the data. The corresponding ROC curves, in figure 3, show that the change in the class definition in the FC data produces little to no effects on the ranking performance of the Naive Bayes classifier. In

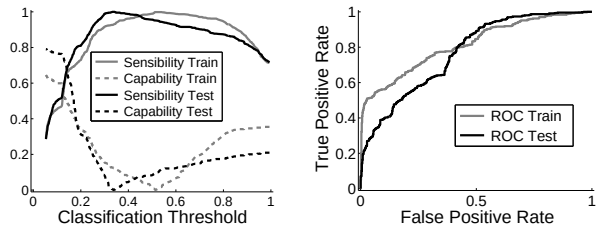

Figure 2. Left: Sensibility and Capability curves. Right: the ROC Curve for training NGP and testing DP domains.

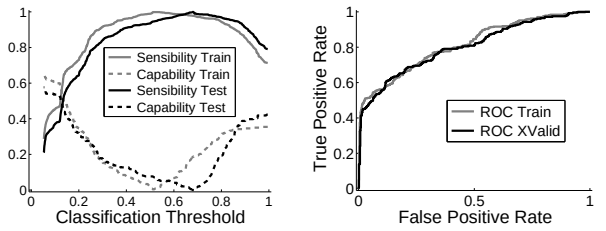

Figure 3. Left: Sensibility and Capability curves. Right: the ROC Curve for training NGP and testing FC domains.

fact, a lower struggle ratio also confirms an improved probability estimation on the FC test data. Never-theless, the test data produces sensibility and capability curves that are visually different from those obtained for training on the original domain, the NGP domain. Therefore, we argue that the ROC analysis fails to depict changes in the underlying distribution which do not impact the classification performance. The use of probability scores in our analysis remedies this situation by depicting classification sensibility.

4. Conclusion

In this paper, we argue that the ROC analysis depicts the ranking performance of a probabilistic classifier independent of its probability estimation performance. Since the probabilities are assigned to examples by the classifier, they are also eliminated from the process of performance analysis, therefore, the ROC curve may be unable to detect changes in the distribution of these probabilities. The ROC analysis remains consistent with the assumption that training and testing data are drawn from the same static distribution. However, this assumption is criticized in practical applications. In fact, it may be desired to include such changes in the performance analysis. To this extent, we propose the method of sensibility analysis to measure changes in the underlying distribution. Our method utilizes class probabilities to determine classification sensibility. Further, the use of class probabilities as part of the evaluation process supports our Soft ROC analysis presented in the 2008 workshop (Klement & Flach, 2008).

Several extensions may be applied to our method of classification sensibility. Our metrics can be used to assess the quality of time-stamped data, such that, when a change in the domain occurs, our method can determine whether a particular test set is suitable for evaluating a given learning method. In addition, it may be interesting to study the effect of optimizing classification sensibility during the learning phase. Such studies remain in our future work.

5. Acknowledgement

We thank Peter Flach for his contribution in the way the midpoint t_s is calculated. We thank the Natural Sciences and Engineering Research Council of Canada (NSERC) and the Ontario Centres of Excellence (OCE) for financial support.

References

Alaíz-Rodríguez, R., & Japkowicz, N. (2008). Assessing the impact of changing environments on classifier performance. *In proc. of Canadian AI'08* (pp. 13–24).

Hand, D. (2006). Classifier technology and the illusion of progress. *Statistical Sciences*, *21*, 1–15.

Klement, W., & Flach, P. (2008). Soft receiver operating characteristics curves. *In proc. of the ICML 2008 Workshop on Evaluaiton Methods for ML.*.

Klinkenberg, R. (2004). Learning drifting concepts: Example selection vs. example weighting. *Intelligent Data Analysis*, *8*, 282–300.

Narasimhamurthy, A., & Kuncheva, L. (2007). A framework for generating data to simulate changing environments. *In proc. of the IASTED Int. Conf. on AI and Applications* (pp. 384–389).

Provost, F., & Fawcett, T. (2001). Robust classification systems for imprecise environments. *Machine Learning*, *42*, 203–231.

Widmer, G., & Kubat, M. (1996). Learning in the presence of concept drift and hidden contexts. *Machine Learning*, *32*, 69–101.