Accepted in Int. Conference on Pattern Recognition (ICPR), 2012

Adaptive Selection of Ensembles for Imbalanced Class Distributions

Paulo V. W. Radtke[1], Eric Granger[1], Robert Sabourin[1], Dmitry Gorodnichy[2]
[1]École de technologie supérieure, Université du Québec, Montreal, Canada
[2]Science and Engineering Directorate, Canada Border Services Agency, Ottawa, Canada
radtke@livia.etsmtl.ca, {Eric.Granger, Robert.Sabourin}@etsmtl.ca, dmitry.gorodnichy@cbsa-asfc.gc.ca

Abstract

Boolean combination (BC) techniques have been shown to efficiently integrate the responses of multiple classifiers in the ROC space for improved accuracy and reliability. Although the impact on classification performance of imbalanced class distributions may be addressed using ensemble-based techniques, it is difficult to observe with ROC curves. Given a false alarm rate and class imbalance, performing BC in the Precision-Recall Operating Characteristic (PROC) space can lead to a higher level of performance. In practice, class distributions often change over time, and BCs should adapt to reflect operational conditions. Thus, this paper proposes an adaptive system that initially uses skewed data to generate several BCs in the PROC space. Then, during operations, the class imbalance is periodically estimated, and used to estimate the most accurate BC of classifiers among operational points of these curves. Simulation results indicate that this approach maintains a level of accuracy that is comparable to full Boolean re-combination, but for a significantly lower computational cost.

1 Introduction

In several real world classification applications, class distributions are imbalanced and change over time. For instance, in public sector video surveillance, face recognition across a network of IP cameras allow an enhanced screening of individuals of interest in dense and moving crowds. One specific challenge in video surveillance is that only a small proportion of the faces captured during operations correspond to an individual of interest. Neural and statistical classifiers for face matching are typically trained on balanced data to maximize accuracy, although actual class priors are unknown a priori and may change over time.

Four main approaches to deal with class imbalance are discussed in the literature [4, 11]. At the **algorithm level**, the learner behavior is modified to bias toward the minority (positive) class. It is difficult to achieve as it depends on understanding both the learner principles and the problem, such as scaling priors on MLP neural networks. A **cost sensitive approach** changes the learning procedure to minimize the cost of misclassified instances, where each error type has a different cost (usually higher to the minority class). The drawback with this method is that the missclassification costs are difficult to estimate (problem dependent). *Data level approaches* require no modification to the learner algorithm, and are categorized either as *undersampling* or as *oversampling* techniques. The literature suggest that the accuracy and reliability of a classification system can be improved by integrating the evidence from multiple different sources of information [6]. **Ensemble of classifiers** (EoCs) have been used [4, 7] to address classification problems with imbalanced class distributions, with the advantage that they do not necessarily require changes to base classifiers.

Boolean combination (BC) techniques [9] can efficiently combine the decisions of several crisp or soft 1- or 2-class classifiers, optimizing the combination of decision thresholds (operation points) with respect to performance. In literature, BC is usually performed in the Receiver Operating Characteristics (ROC) space [3]. However, since the impact on performance of class imbalances is difficult to observe with ROC curves, performing BC in the *Precision-Recall Operating Characteristic* (PROC) [10] space may provide a higher level of performance. Estimating the precision (in conjunction with recall) is more appropriate in imbalanced settings, as it remains sensitive to the performance on each class. Moreover, since class distributions are subjected to continuous or seasonal changes, the BC should be adapted over time to reflect new operational conditions.

In this paper, an adaptive system is proposed to select the BC of classifiers given the desired false alarm rate (far) and class imbalance. During design phases,

skewed validation data is used to generate several BCs in the PROC space, by successively growing number of samples from the majority class. Then, during operations, the system relies on the Hellinger distance to periodically detect and estimate changes to class distributions from operational data streams. Once a change has been detected, class imbalance is estimated, and the closest operational points on PROC curves are employed to estimate the combination of classifiers.

2 Boolean Combination

A soft classifier C produces a binary decision when its response is compared to a threshold $\gamma \in [0, 1]$. For a threshold set Γ, classifier C has the operation points $C_\gamma, \gamma \in \Gamma$, providing a performance tradeoff between classes. BC of two soft classifiers C_a and C_b is the fusion of all $C_{a,\gamma}$ and $C_{b,\gamma}$ through Boolean operations, thus, each resulting operation point is an EoC based on thresholds and a Boolean function. Selecting the non-dominated operational points in the decision space (for instance, the ROC convex hull) defines the operation points with the best trade offs. Haker *et al.* [8] used a set of Boolean operators with the conjunction and disjunction ($O = \{\vee, \wedge\}$). Khreich *et al.* [9] used an iterative variation with 10 Boolean operations. Both works operate in the ROC space.

The *Receiver Operator Characteristics* (ROC) [3] analysis is based on two intra-class measures, the *true positive rate* tpr (proportion of correct positive class predictions) and the *false positive rate* fpr (proportion of incorrect negative class predictions). BC in the ROC space select operation points in the convex hull (best trade off between tpr and fpr), where each vertex is a Boolean fusion of classifiers. However, ROC analysis is insensitive to class imbalances. The *Precision-Recall Operating Characteristic* (PROC) [10] analysis focus on an inter class measure, classifier *precision* (proportion of correct positive predictions), and *recall*, the same as tpr. Thus, PROC graph plots represents classifier performance regarding data skew. It is demonstrated in [1] that operating points that belong to the ROC convex hull also belong to the PROC achievable curve. Thus, one can find those operating points in the ROC space to perform BC, and compare them in the PROC space to consider different data skew levels when comparing BCs. Based on experimental results, selecting and adapting combinations in the PROC space is better for imbalanced data.

3 Adaptive Selection of Ensembles

Figure 1 shows the architecture of an adaptive system to select the most accurate BC of classifiers according to class imbalance. Assume a pool of detectors $D_1, \ldots, D_n$, each connected to a sensor. Given a signal $\mathbf{x}_i(t)$ captured by sensor i in time t, detector i extracts and selects features, providing a feature vector $\mathbf{f}_i$ to a 1- or 2-class classifier C_i. Continuous scores $s_i(\mathbf{x}_i) \in [0, 1]$ are compared against thresholds values in Γ through BC to provide an overall decision $d(t)$.

During design, BC is performed in the PROC space according to several levels of class imbalance, by successively growing the number of samples from the majority (negative) class w.r.t. the ones in the positive class. One level of imbalance is selected a priori for operations. To maintain accuracy over time, the BC requires periodical adaptation to reflect current class imbalances based on samples captured during operation. The adaptation module is used at application-dependent intervals to estimate the class imbalance, and update the BC (Boolean fusion and threshold values). The Hellinger distance [5] is used to estimate changes to class distributions from operational data streams. Once a change has been detected, class imbalance is estimated and the most accurate combination of classifiers is estimated based on the closest operational points on PROC curves.

Class imbalance relates to data skew, λ, the ratio of positive samples π_p to negative ones π_n. A $\lambda = 0.01$ indicates that for each positive sample we have 100 negative samples, or $\pi_p : \pi_n = 1 : 100$. In literature, some approaches have been proposed to estimate class imbalance, and these are useful to adapt classifier combinations with BC. Online approaches may be either based on transductive learning [12] or transfer learning [13]. The Hellinger distance [2, 5] allows to estimate the imbalance between a distribution of unlabeled operational data and of labeled training data for BC. In this paper, operational class proportions are periodically estimated against validation data. Given a set of operational data

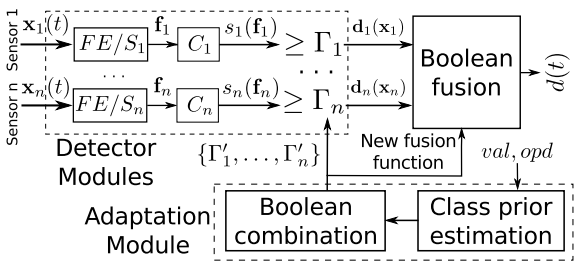


Figure 1: Architecture to adapt a BC of classifiers to imbalanced class distributions.

opd and the validation data val, the Hellinger distance $H(val, opd)$ at the feature level (each sample with n_f features) is calculated for each feature f using discrete distributions (bins) with a probability associated to each bin b in the feature space using

$$H(val, opd) = \frac{1}{n_f} \sum_{f=1}^{n_f} \sqrt{\sum_{i=1}^{b} \left(\sqrt{\frac{|val_{f,i}|}{|val|}} - \sqrt{\frac{|opd_{f,i}|}{|opd|}} \right)^2} \quad (1)$$

where $|val_{f,i}|$ is the number of samples in val that for feature f are within bin b limits. The same applies to opd with $|opd_{f,i}|$. Given normalized features in the $[0, 1]$ interval and limited positive data for training, a set of $b = 20$ bins is assumed.

As the class proportions in val and opd are closer, $H(val, opd)$ tends towards zero. Equation 1 helps to build a new labeled val^* validation data set to perform BC with the correct class proportions. Skew is then calculated as $\lambda = \pi_p/\pi_n$. To avoid the costly full recomputation of BCs for each new λ, Algorithm 1 is proposed to approximate the fusion function. Every time classifiers are combined during design for a λ value, a set of operational points are generated and stored (set of ensembles E), each tagged with the appropriate data skew level. When a new data skew λ^* is detected and the set of ensembles for data skews $1 : i$ and $1 : k$, $i < j < k$, are available, the BC is approximated. That is, the outer envelop of the PROC curves for both original skew levels, $1 : i$ and $1 : k$, are combined using val^* data with the newly detected λ^*. The resulting EoC p is then used to update the BC in Fig. 1.

Data: Classifiers set C, thresholds set Γ, data set val^* with skew λ^*, set E of optimized BCs for different skew levels (each a set of ensembles, one for each operational point), the desired false alarm rate far
Result: Ensemble p and the updated set E
$F = \emptyset$;
if $\exists E^*_\lambda \in E$, *with skew level* λ^* **then**
  $F = E_{\lambda^*}$;
end
if $\exists E_{\lambda^1}, E_{\lambda^2} \in E$ *with skew levels* λ^1, λ^2, *such as that* $\lambda^1 < \lambda^* < \lambda^2$ **then**
  $E' = E_{\lambda^1} \cup E_{\lambda^2}$;
  forall $d \in E'$ **do**
    $F = F \cup d$ iff $\nexists e \in E', e \succ d$ on val^*;
  end
else
  Obtain Boolean combination BC with for C, Γ, and val^*;
  $E = E \cup \{BC\}$;
end
Select EOC (operational point) $p \in BC$ at the desired far;

Algorithm 1: Adapting BC for class imbalance.

Skew	Balanced validation		Skewed validation	
	accuracy	F-Measure	accuracy	F-Measure
1:1	78.40%	0.743	78.40%	0.743
1:10	91.43%	0.569	91.90%	0.577
1:100	94.21%	0.176	94.66%	0.183
1:1000	94.50%	0.022	94.97%	0.024

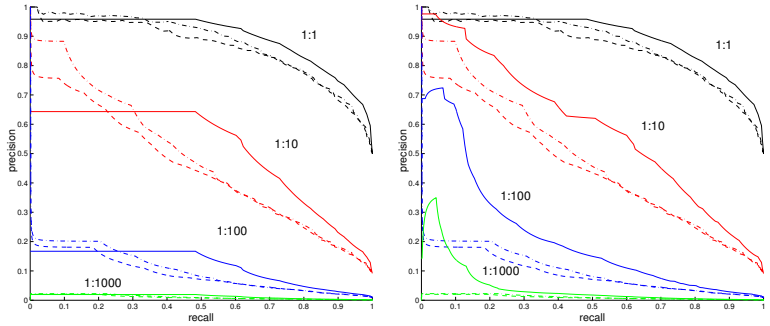
Table 1: BCs at 5% far on imbalanced test data.

4 Simulation Results

Two proof-of-concept experiments are performed using synthetic bi-dimensional data, Gaussian distributions centered at $(0, 0)$ (positive class) and $(1.5, 1.5)$ (negative class) using the identity matrix as the covariance matrix. The training data is used to train two different linear discriminant classifiers (LDC), C_1, trained with the abscissa sample values, and C_2, trained with the ordinate sample values. In these experiments, the BC technique of Haker et al. [8] is used with $O = \{\vee, \wedge\}$.

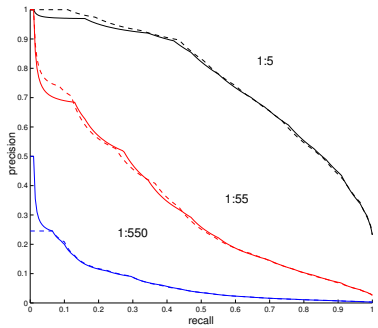
In **Experiment 1**, the impact on BC of imbalanced data is observed with C_1 and C_2. As a first step, LDCs are trained with 100 samples per class, and BC is performed with a balanced val data set of 200 samples per class. Resulting EoCs are tested against a $test$ data set with 1000 positive samples and different proportions of negative samples, $1 : 10^n, 0 \leq n \leq 3$. In the second step, the val set for BC uses the same skews as $test$. Results indicate that changing the level of skew in val data also changes the resulting BC EoCs. However, this change is difficult to observe in the ROC space – ROC curves and scalar AUC measures are equivalent. Indeed, ROC analysis is insensitive to class imbalance since both fpr and tpr are intra class measures. Plotting these EoCs in the PROC space in Figs. 2a and 2b show the impact of data skew in val during BC for problems with imbalanced classes. For each level of skew, PROC curves detail the performance for LDCs alone and the curve obtained through BC. It is first observed that BC using imbalanced val covers a large area in the plot, providing better trade offs. Selecting an operational point, e.g., $far = 5\%$ further supports the use of skewed validation data, as accuracy and F_1 scores are consistently higher (see Table 1). The accuracy increase for a same skew level, and the $far = 5\%$ translates to an improvement of the positive class prediction and improvement on the F_1 scores.

Experiment 2 validates Alg. 1 (adaptation module in Fig. 1). Assuming the detection of three different in-

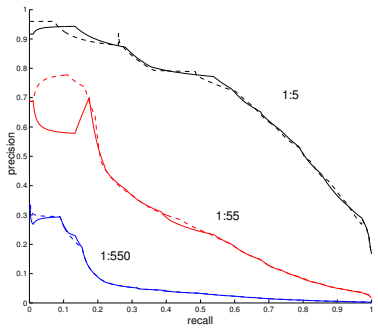


(a) BC on test using balanced val data.

(b) BC on test using imbalanced val data.



(c) Approximated curves on val data.



(d) Approximated curves on $test$ data.

Figure 2: BC PROC curves. In 2a and 2b, solid lines are the BC, dashed lines c_1 and c_2. In 2c and 2d, solid lines are the BC approximation and dashed lines the actual BC.

termediate class imbalances, $1:5$, $1:55$ and $1:550$, new BCs are approximated using the BCs optimized in Experiment 1. For comparison, the actual BC for these skew levels are also calculated. Fig. 2c (validation) and Fig. 2d (test) compares the PROC curves of actual and approximated BCs. Curves for each skew level are equivalent, the same for precision and F_1 scores for a 5% far in Table 2. Algorithm 1 is also computationally more efficient. For $n = 2$ classifiers, a traditional BC requires $|T|^n \times |op| + |T| \times n$ operations to evaluate the tpr and fpr values. For the simulations in this paper with $|T| = 100$, a total of 20200 evaluations are required. Approximating with Algorithm 1 requires $(|H_i| + |H_k|) \times (|op| + n)$, which averaged to 184 evaluations, a significant reduction on computational effort.

Skew	Actual combination		Approximated combination	
	accuracy	F-Measure	accuracy	F-Measure
1:5	89.37%	0.658	89.30%	0.654
1:55	94.40%	0.281	94.40%	0.281
1:550	94.94%	0.042	94.94%	0.042

Table 2: Approximated BCs at 5% far on test.

5 Discussion

EoCs have been proposed in the literature to reduce the impact from imbalanced class distributions. BC of ensembles on the ROC space have been shown to improve accuracy and reliability, although the impact of imbalanced class proportions is difficult to observe with ROC curves. Experiments in this paper show that performing BC in the PROC space produces a better combination of base classifiers. In this paper, an adaptive system is proposed to select the most accurate BCs according to the desired far and class imbalance. Skewed validation data is used to generate several BCs with PROC curves, by successively growing number of samples from the majority class. During operations, the system periodically detects changes to class proportions from operational data, and estimated class imbalance. The closest operational points on PROC curves are employed to estimate the most accurate BC of classifiers. Instead of full BC, the knowledge obtained when combining classifiers for other skew levels is used to approximate the BC to new class priors, providing a significant reduction in computational complexity.

References

[1] J. Davis and M. Goadrich. The relationship between precision-recall and roc curves. In *Proceedings of the 23rd international conference on Machine learning*, ICML '06, pages 233–240, New York, NY, USA, 2006. ACM.

[2] G. Ditzler and R. Polikar. Hellinger distance based drift detection for nonstationary environments. In *CIDUE Symposium*, 2011.

[3] T. Fawcett. An introduction to roc analysis. *Pattern Recogn. Lett.*, 27:861–874, June 2006.

[4] M. Galar, A. Fernandez, E. Barrenechea, H. Bustince, and F. Herrera. A review on ensembles for the class imbalance problem: Bagging-, boosting-, and hybrid-based approaches. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, PP(99):1–22, 2011.

[5] V. González-Castro, R. Alaiz-Rodríguez, L. Fernández-Robles, R. Guzmán-Martínez, and E. Alegre. Estimating class proportions in boar semen analysis using the hellinger distance. In *Proceedings of the 23rd international conference on Industrial engineering and other applications of applied intelligent systems - Volume Part I*, IEA/AIE'10, pages 284–293, Berlin, Heidelberg, 2010. Springer-Verlag.

[6] E. Granger, W. Khreich, and D. O. Gorodnichy. Fusion of biometric systems using boolean combination: An application to iris-based authentication. *International Journal on Biometrics, in press*.

[7] X. Guo, Y. Yin, C. Dong, G. Yang, and G. Zhou. On the class imbalance problem. In *Natural Computation, 2008. ICNC '08. Fourth International Conference on*, pages 192–201, 2008.

[8] S. Haker, W. W. III, S. Warfield, I.-F. Talos, J. Bhagwat, D. Goldberg-Zimring, A. Mian, L. Ohno-Machado, and K. Zou. Combining classifiers using their receiver operating characteristics and maximum likelihood estimation. 8(Pt 1):506–14, 10 2005.

[9] W. Khreich, E. Granger, A. Miri, and R. Sabourin. Iterative boolean combination of classifiers in the roc space: An application to anomaly detection with hmms. *Pattern Recognition*, 43(8):2732 – 2752, 2010.

[10] T. Landgrebe, P. Paclik, R. Duin, and A. Bradley. Precision-recall operating characteristic (p-roc) curves in imprecise environments. In *Pattern Recognition, 2006. ICPR 2006. 18th International Conference on*, volume 4, pages 123 –127, 0-0 2006.

[11] W.-J. Lin and J. J. Chen. Class-imbalanced classifiers for high-dimensional data. *Briefings in Bioinformatics Advance*, 2012.

[12] C. Yang and J. Zhou. Non-stationary data sequence classification using online class priors estimation. *Pattern Recogn.*, 41(8):2656–2664, Aug. 2008.

[13] Z. Zhang and J. Zhou. Transfer estimation of evolving class priors in data stream classification. *Pattern Recogn.*, 43(9):3151–3161, Sept. 2010.