Adaptive Selection of Ensembles for Imbalanced Class Distributions

Paulo V. W. Radtke[1], Eric Granger[1], Robert Sabourin[1] and Dmitry Gorodnichy[2]
[1]Laboratoire d'imagerie, de vision et d'intelligence artificielle – LIVIA
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 diversified classifiers in the ROC space to improve the overall accuracy and reliability of pattern recognition systems. In practice, since class distributions are often imbalanced and change over time, the BC of classifiers, and thus selection of ensembles, should be adapted to reflect operational conditions. Although the impact on classification performance of imbalanced distributions may be addressed using ensemble-based techniques, this is difficult to observe from ROC curves. However, given a desired false positive rate and class imbalance, performing BC in the Precision-Recall Operating Characteristic (PROC) space with skewed data may lead to a higher level of performance. In this paper, an adaptive system is proposed that initially generates several PROC curves, each one from data with a different level of skew. Then, during operations, the class imbalance is periodically estimated, and used to approximate the most accurate BC of classifiers among operational points of these curves. Simulation results indicate that this approach maintains a high level of accuracy that is comparable to full Boolean re-combination (as required for a specific level of imbalance), but for a significantly lower computational cost.

1 Introduction

In several real-world pattern recognition applications, underlying class distributions are imbalanced and change over time. For instance, in public sector video surveillance, face recognition across a network of IP cameras allows for enhanced screening of individuals of interest in cluttered and moving crowds. In decisions support systems, an human operator employ video face recognition to track and recognize the facial regions captured across frames, and attempt to detect individuals of interest. Class distributions may change gradually, recurrently and abruptly according to capture conditions (density of people in the scene, illumination, etc.) and physiology (ageing).

Another specific challenge in video surveillance is that only a small proportion of the faces captured during operations correspond to an individual of interest. One- or two-class neural and statistical classifiers for face matching are typically trained using a balanced data to avoid biasing performance toward the majority class, and then classifier outputs are scaled according to prior knowledge. Although training samples are abundant for the negative class (non-target individuals), positive class samples are typically limited, and the proportion of positive to negative training samples is not consistent with operational data and unknown class priors.

Some approaches have been used to estimate changes in class distributions during operation [12, 13, 2, 5], and four main approaches have been proposed for training classifiers with class imbalance in the literature [4, 11]. At the **algorithm level**, the learner behavior is modified to bias toward the minority (positive) class. One example is to scale 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 and problem dependent). **Data level approaches** require no modification to the learner algorithm, and are categorized either as undersampling or as oversampling techniques. Finally, **ensemble of classifiers** (EoCs) have been used [4, 7] for learning on classification problems with imbalanced class distributions, with no required changes to base classifiers.

Besides addressing class imbalance, the literature also suggests that the accuracy and reliability of a classification system can be improved by integrating the

evidence from multiple different sources of information [6]. 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 (operational points) with respect to performance. In the decision space, each vertex represents an operational EoC, and virtual EoCs are obtained through linear interpolation between two vertexes. Given a class imbalance, it is possible to produce a decision-level fusion function and thresholds through BC.

In the literature, BC is usually performed in the *Receiver Operating Characteristics* (ROC) space [3]. However, since the impact on accuracy of class imbalances is difficult to observe with ROC curves, performing BC with imbalanced data in the *Precision-Recall Operating Characteristic* (PROC) [10] space may provide a higher level of accuracy for the same false positive rate. Estimating the precision (in conjunction with recall) is more appropriate in imbalanced settings, as it remains sensitive to the performance on both classes.

This paper proposes a technique to adapt the selection of BCs for a desired false positive rate in response to changing levels of class imbalance. The main difference to classical approaches is to detect class imbalance over operational data to adapt an ensemble-based system for the target operational false positive rate, with a low computational cost in comparison to full BC. 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 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 approximate the combination of classifiers. This process should reduce the computational time w.r.t. full BC.

The remainder of this paper is organized as follows. Section 2 discusses Boolean combination in the decision space, and Section 3 proposes the approach to adapt the BC of EoCs for imbalanced data. Section 4 details experiments with synthetic data and Section 5 discusses the results obtained.

2 Boolean Combination

A soft classifier C produces a binary decision when its response is compared to a threshold $\gamma, 0 \leq \gamma \leq 1$. For a set of thresholds $\Gamma = \{\gamma_1, \ldots, \gamma_m\}$, the classifier C has the operational points $C_\gamma, \gamma \in \Gamma$, providing a performance trade off 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. Therefore, each resulting operational point is an EoC based on thresholds and a Boolean fusion 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 only fusion functions with the Boolean conjunction and disjunction operators ($\vee$ and $\wedge$), under the assumption of conditionally independent classifiers. Khreich *et al.* [9] proposed a more general iterative variation with 10 Boolean fusion operations. Both works operate in the ROC decision 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 operational 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) along with *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 and fpr. 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 $d_i(t_i)$.

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 re-

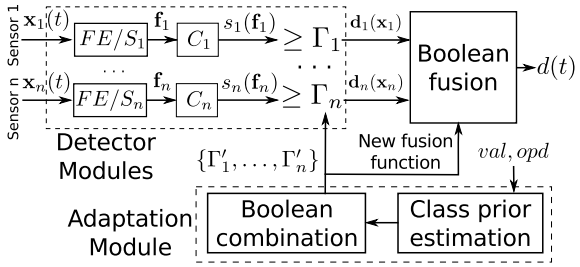


Figure 1: Architecture to adapt a BC of classifiers to imbalanced class distributions.

quires 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 detect 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 approximated, based on the closest operational points on PROC curves.

In literature, some approaches have been proposed to estimate and detect changes in class imbalance, or skew [1] λ, 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 detect the imbalance between a distribution of unlabeled operational data and of labeled training data for BC. In this paper, operational class proportions changes are periodically detected against validation data. Given a set of operational data 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 b bins in the feature space

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

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 $|opd_{f,i}|$.

As the class proportions in val and opd are closer, $H(val, opd)$ tends towards zero. Equation 1 detects class imbalance changes, and the class imbalance is estimated to build a new labeled val^* validation data set

[1] Class imbalance relates to data skew, λ, the ratio of positive samples π_p to negative ones π_n. Thus, $\lambda = \pi_p/\pi_n$ and $\lambda = 0.01$ indicates that for each positive sample we have 100 negative samples, or $\pi_p : \pi_n = 1 : 100$.

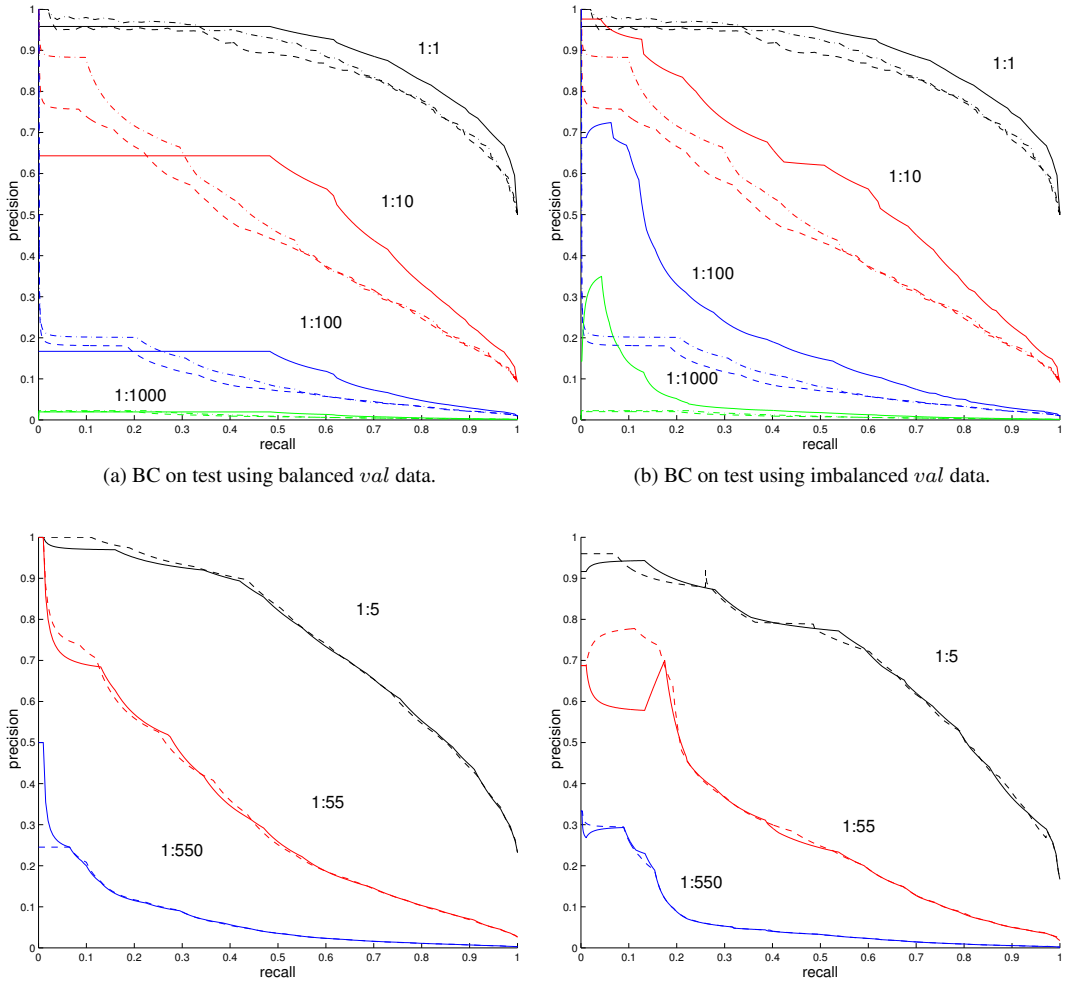
Data: Classifiers set C, thresholds set Γ, data set val^* with skew λ^*, set E of optimized BCs for different skew levels (each a set of ensembles, one for each operational point), the desired false positive rate fpr
Result: Ensemble p and the updated set E
$F = \emptyset$;
if $\exists E_{\lambda^*} \in E$, *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 the $d \in E'$ **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 fpr;

Algorithm 1: Adapting BC for class imbalance.

to perform BC with the correct class proportions. To avoid the costly full re-computation of BCs for each new value of λ, Algorithm 1 is proposed to approximate the Boolean 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 $\lambda^* = 1 : j$ is detected and the set of ensembles for data skews $\lambda^1 = 1 : i$ and $\lambda^2 = 1 : k$, $i < j < k$, are available, the BC is approximated. That is, the outer envelop of the PROC curves for both original skew levels, λ^1 and λ^2, are combined using val^* data with the newly detected λ^*. The resulting EoC p is then used to update the BC in Fig. 1.

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 proposed by Haker et al. [8] is used with $O = \{\vee, \wedge\}$.



(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.

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 operational trade offs. Selecting an operational point, e.g., $fpr = 5\%$ further supports the use of skewed validation data, as accuracy and F_1 scores are consistently higher (see Table 1). The accuracy increase for the same skew level and the same $fpr = 5\%$ 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 intermediate 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 λ	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: Performance on test data at $fpr = 5\%$ of BCs obtained with balanced and imbalanced validation data.

Skew λ	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: Performance on test data at $fpr = 5\%$ of actual and approximated BCs obtained with Algorithm 1.

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 $fpr = 5\%$ 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.

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 fpr 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 estimates class imbalance. The closest operational points on PROC curves are employed to approximate 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 for real time operation.

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*, pages 233–240, New York, NY, USA, 2006.

[2] G. Ditzler and R. Polikar. Hellinger Distance Based Drift Detection for Nonstationary Environments. In *IEEE Symposium on Computational Intelligence in Dynamic and Uncertain Environments*, pages 41–48, 2011.

[3] T. Fawcett. An Introduction to ROC Analysis. *Pattern Recognition Letters*, 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*, 42(4):463–484, 2012.

[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*, pages 284–293, 2010.

[6] E. Granger, W. Khreich, R. Sabourin, and D. O. Gorodnichy. Fusion of Biometric Systems Using Boolean Combination: An Application to Iris-Based Authentication. *International Journal on Biometrics*, 4(3):291–315, 2012.

[7] X. Guo, Y. Yin, C. Dong, G. Yang, and G. Zhou. On the Class Imbalance Problem. In *Proceedings of the 4th International Conference on Natural Computation*, pages 192–201, 2008.

[8] S. Haker, W. W. III, et al. Combining Classifiers Using their Receiver Operating Characteristics and Maximum Likelihood Estimation. In *Proceedings of the 8th International Conference on Medical Image Computing and Computer Assisted Interventions*, pages 506–514, 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 *18th International Conference on Pattern Recognition*, pages 123–127, 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 Recognition Letters*, 41(8):2656–2664, Aug. 2008.

[13] Z. Zhang and J. Zhou. Transfer Estimation of Evolving Class Priors in Data Stream Classification. *Pattern Recognition Letters*, 43(9):3151–3161, Sept. 2010.