Information Fusion 20 (2014) 31–48

Contents lists available at ScienceDirect

Information Fusion

journal homepage: www.elsevier.com/locate/inffus

Skew-sensitive boolean combination for adaptive ensembles – An application to face recognition in video surveillance

Paulo V.W. Radtke [a,*], Eric Granger [a], Robert Sabourin [a], Dmitry O. Gorodnichy [b]

[a] Laboratoire d'imagerie, de vision et d'intelligence artificielle, École de technologie supérieure, Université du Québec, Montreal, Canada
[b] Science and Engineering Directorate, Canada Border Services Agency, Ottawa, Canada

ARTICLE INFO

Article history:
Available online 21 November 2013

Keywords:
Classifier ensembles
Class imbalance
Information fusion
Boolean combination
Face recognition

ABSTRACT

Several ensemble-based techniques have been proposed to design pattern recognition systems when data has imbalanced class distributions, although class proportions may change over time according to the operational environment. For instance, in video surveillance applications, face recognition (FR) is employed to detect the presence of target individuals of interest in potentially complex and changing environments. Systems for FR in video surveillance are typically designed a priori with a limited amount of reference target data and prior knowledge of underlying class distributions. However, the relatively proportion of target and non-target faces captured during operations varies over time. Estimating the actual proportion of data from the input data stream could allow to dynamically adapt ensembles to reflect operational conditions. In this paper, the selection and fusion of ensembles produced through Boolean Combination (BC) of classifiers is periodically adapted based on the class proportions estimated from input streams. BC techniques have been shown to efficiently integrate the responses of multiple diversified classifiers in the ROC space, yet the impact on performance of imbalanced data distributions is difficult to observe from ROC curves. Given a diversified pool of classifiers and a desired false positive rate (*fpr*), the new Skew-Sensitive Boolean Combination (SSBC) technique exploits the Precision-Recall Operating Characteristic (PROC) space, leading to a higher level of performance. A set of BCs of base classifiers is initially produced with imbalanced reference data in the PROC space, where each BC curve corresponds to different level of imbalance (a growing number of non-target samples versus a fixed number of target ones). Then, during operations, the closest adjacent levels of class imbalance are periodically estimated using the Hellinger distance between the data distribution of inputs and that of imbalance levels, and used to approximate the most accurate BC of classifiers from operational points of these curves. Simulation results on Faces In Action video surveillance data indicate that ensemble-based FR systems using the SSBC technique outperform the same systems using traditional BC techniques with Random Under-Sampling and One-Sided Selection. It allows to dynamically select BCs that provide a higher level of precision (and F1 value) for target individuals, and a significantly smaller difference between desired and actual *fpr*. Performance of this adaptive approach is also comparable to the costly full recalculation of BCs (as required by a BC technique to accommodate a specific level of imbalance), but for a computational complexity that is considerably lower. Finally, SSBC is shown to achieve a high level of discrimination between target and non-target individuals when face tracking is exploited to accumulate ensemble predictions for facial captures that correspond to a same person in the video scene.

1. Introduction

Issues related to class imbalance are present in many real-world applications, and several algorithm-level, data-level, cost-sensitive and ensemble-based techniques have been developed to design pattern recognition systems when data has imbalanced class distributions [1]. Since the proportion of design samples per class rarely correspond to the actual distribution of operational data, system performance may differ from that achieved during the design phase. What is more, the underlying class distributions change over time in many applications. It is therefore desirable to estimate class imbalance over time, and adaptively select an operational point with design data that follows the class imbalance of operational data.

As a practical example, consider systems for automated face recognition (FR) in video surveillance. Theses systems allow to rapidly detect and track individual of interest across a network of

* Corresponding author.
E-mail addresses: radtke@livia.etsmtl.ca (P.V.W. Radtke), Eric.Granger@etsmtl.ca (E. Granger), Robert.Sabourin@etsmtl.ca (R. Sabourin), dmitry.gorodnichy@cbsa-asfc.gc.ca (D.O. Gorodnichy).

1566-2535/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.inffus.2013.11.001

video surveillance cameras for enhanced security and situational awareness. Given an individual or situation of interest, the operator of a human-centric decision support system may, for instance, capture reference samples (facial regions) corresponding to an individual appearing in video feeds, and then store a facial model, e.g., a template or statistical representation, designed using theses samples. Face re-identification consists in matching facial regions captured in live or archived video streams against facial models of individuals enrolled to a system [2–4].

Face re-identification is typically performed in either semi-constrained (e.g., an inspection lane) or unconstrained (e.g., portals or free-flow airport scenes) environments, which raises several challenges. Systems for FR are exposed to complex operational environments with changing illumination, pose, expression, occlusion, etc. In addition, a relatively small proportion of the faces collected for system design or captured during operation correspond to an individual of interest, although non-target faces may appear in abundance. The covert and unobtrusive capture of video sequences with target individuals provides only a limited amount of high quality reference samples to design facial models. To avoid biasing performance towards the majority class, neural and statistical classifiers used for face matching are typically designed with balanced data, provided by using data-level sampling techniques. An estimate of class priors is often used to scale classifier outputs, although actual class proportions are often unknown a priori and change over time.

Several specialized architectures have been proposed for FR in video surveillance [5–8]. Given the need to detect individuals of interest, some systems have effectively been modeled in terms of independent user-specific detectors, each one implemented using one or more binary (1- or 2-class) classifiers [9,5]. For example, this modular approach was employed in [5] with user-specific ensembles of 2-class ARTMAP neural network classifiers and Boolean combination (BC) of responses. However, these architectures do not consider the impact on performance of class imbalance. Exploiting this information to enhance performance is relevant in video surveillance, owing to the potentially small number of target samples w.r.t. the abundant non-target samples.

BC techniques [10] can efficiently integrate the responses of several crisp or soft binary classifiers, optimizing the combination of decision thresholds (operational points) with respect to accuracy or another scalar performance metric related to a decision space. BC is typically performed in the Receiver Operating Characteristics (ROC) space [11]. Once a pool of diversified classifiers has been generated, a BC technique is used for selection and fusion of ensembles based on the desired false alarm rate, using balanced validation data. In practice, it would however be beneficial to adapt the BC of ensembles (i.e., Boolean functions and thresholds) according to an estimate of class imbalance during operation.

Although the impact on classification performance of class imbalance cannot be observed with ROC curves, the *Precision-Recall Operating Characteristic* (PROC) [12] space (also known as the Precision-Recall space) is sensitive to imbalance. Indeed, the precision measure allow to observe the proportion of correct target predictions over all positive samples predictions, which typically declines when the proportion of negative samples grow with respect to the positive ones. Given the relationship between ROC and PROC space [13], the PROC space can be exploited to adapt classification systems for changing levels of class imbalance.

In this paper, a new BC technique is proposed to adapt the selection and fusion of classifier ensembles given the desired false positive rate and the current level of class imbalance, as estimated from operational data streams. This technique, called the Skew-Sensitive BC (SSBC) technique, exploits the PROC space. During design phases, a pool of diversified classifiers is generated, and imbalanced validation data is used to produce several BC curves in the decision space, by successively growing the number of negative (non-target) samples from the majority class. Negative samples are assumed to be available in large quantities, while the limited number of positive target samples is assumed to be fixed. Each BC curve is optimized for one specific class imbalance level. During operations, the system relies on the Hellinger distance [14] to periodically estimate the closest levels of class imbalance from the operational data stream, by comparing it to a set of known reference imbalanced data sets. This estimation is used to approximate the most suitable set of vertices from BC curves in the PROC space, among BC curves with the closest levels of imbalance. Knowledge obtained when combining classifiers for pre-select levels of class imbalance is used to approximate BCs for input data. Compared to full Boolean recombination (i.e., re-calculation of BC curves for a specific level of imbalance), the proposed approach provides a reduction in computational complexity, and thus a faster response times during operations.

Proof-of-concept experiments have been performed with real-world videos from the Carnegie Mellon University Face in Action database [15] where participants to be recognized appear in a passport verification scenario. A security checkpoint (inspection lane or portal) may witness peaks in the flow of target and non-target individuals. For validation, the FR system have been realized with a modular architecture composed of an ensemble of 2-class classifiers per individual, each one generated using a dynamic particle swarm optimization (DPSO)-based training strategy [16]. Each classifier ensemble of the system is selected and combined using the new SSBC technique. Performance of this technique is compared to that of the same ensemble where BC is optimized with data obtained using Random Under-Sampling (RUS) and One-Sided Selection (OSS) [17]. The level of class imbalance has been varied over time during operations. Transactional-based analysis measures the ROC and PROC performance for systems predictions produced in response to faces captured in videos. Time-based analysis exploits face tracking to regroup facial captures that correspond to a same person in the scene. Performance is based on the accumulation the positive system predictions for high confidence face track, over a fixed window of time. Finally, each individual enrolled to the system is categorized according to the Doddington zoo (subject-based) analysis [18,19].

This paper extends the authors' preliminary research in [20] by presenting an in-depth review and analysis on the design of pattern recognition systems when classes imbalance varies during operations. The SSBC technique is proposed to efficiently adapt the selection and fusion of ensembles according to class proportions. While [20] is limited to the strategy to approximate BCs for intermediate skew levels using synthetic data, the complete system presented in this paper includes mechanisms to track class imbalance over time from operational data streams, and to estimate the reference data sets with the closest level of class imbalance, as needed to adapt BCs for a desired *fpr*. SSBC is applied to the adaptation of ensembles for face re-identification, and is validated using real-world video data. Finally, SSBC is compared to BC techniques that exploitee data-level techniques the handle class imbalance.

This paper is organized as follows. In Section 2, strategies for the design and evaluation of pattern recognition systems under class imbalance are provided. Section 3 reviews techniques for BC of binary classifiers, while Section 4 presents the new SSBC that allows to adapt the selection and fusion of ensembles according to estimated levels of class imbalance. The experimental methodology for proof-of-concept experiments is presented in Section 5, and simulation results are shown and interpreted in Section 6.

2. Binary classification from class imbalance

A common assumption in pattern recognition (PR) literature is that class priors are known and that data distributions are balanced, i.e., instances of all classes are assumed to be equally present in both training and operational data. Real world problems rarely follow this ideal case – class priors are unknown and may change over time, and training samples are imbalanced and are not necessarily representative of operational data. For instance, FR in video surveillance involves class imbalance between targets (positive class) and non-targets (negative class) individuals.

Fig. 1 details a two class classification problem, with the unknown underlying distribution indicated for both classes in Fig. 1a. Classifier design seeks to estimate models that generalizes well during operations, using reference data. Fig. 1 illustrates a histogram representation of the distribution obtained from the limited data used for classifier design. The balanced reference data is randomly sampled from the underlying distribution of Fig. 1a. Finally, Fig. 1c illustrates a model of data distributions estimated from the reference data. Densities are estimated with a unimodal Gaussian distribution per class, $\mathcal{N}_+$ and $\mathcal{N}_-$. Since the true prior probabilities are unknown, we estimate that $\widehat{P}_+ = \pi_+$ and $\widehat{P}_- = \pi_-$, where π_+ and π_- are the proportion of positive and negative reference samples and that $\pi_+ = \pi_-$.

Given the underlying distribution and data sampling process, data distributions in Fig. 1 may be imbalanced, or skewed. Data skew (level of class imbalance) is defined as the ratio of positive samples π_p to negative ones π_n, $\lambda = \pi_p/\pi_n$. A skew of $\lambda = 1/100$ indicates that for each positive sample, there are 100 negative samples.

Design of classifiers for FR in video surveillance must also consider the presence of a majority class to define decision rules. Assume a modular classification architecture adapted for surveillance applications, where each target individual is modeled as a user-specific detection module. Each module is implemented with a binary (1- or 2-class) classifier that is assigned to discriminate between the target (positive) and non-target (negative) classes. Binary classifiers output a crisp decision or a score that is compared to a decision threshold to provide a final crisp decision.

The rest of this section reviews techniques available in the literature to design classifiers, to evaluate performance of binary classifiers, to estimate class imbalance during operations and to select decision thresholds under class imbalance.

2.1. Design of classifiers under class imbalance

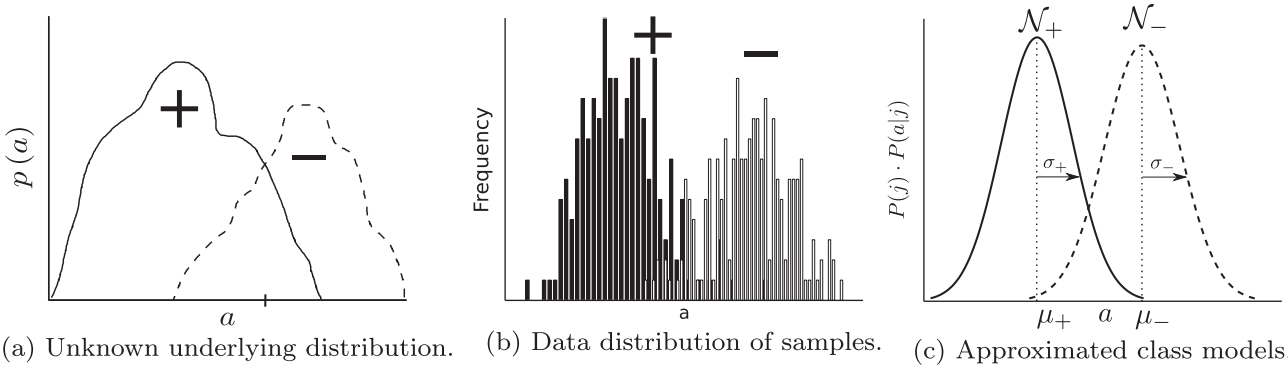
Four main approaches have been proposed in literature to train classifiers from skewed (imbalanced) reference data sets [1,21,22] – algorithm level, cost sensitive, data level and ensemble-based techniques. *Algorithm level approaches* modify the classifier behavior to bias toward the minority (positive) class. Modifying the classifier requires a domain expert that understand the classifier's working principles and the target application. An example is the prior scaling technique in [23], which changes the multi-layer perceptron neural network outputs (or inputs) to better predict the minority class.

Cost sensitive approaches requires fewer modifications (if any) to the classifier. Instead of minimizing the number of misclassified instances, the classifier training procedure minimizes the total cost of misclassified instances. Each error type has a different cost, usually much higher for the minority class, and the sum of miss-classifications costs drives the learning process.

Data level approaches are appealing as they require no modification to the learner algorithm. Data level approaches are categorized either as under-sampling or as over-sampling techniques. Data under-sampling techniques will reduce the sample number of the majority (negative) class to match that of the minority class. Random under-sampling [24] reduces the majority class by randomly selecting instances, which is computationally effective. Other techniques try to follow the majority class distribution in the feature space, such as the Condensed Nearest Neighbor Rule [25] or the One Sided Selection [17], both techniques that emphasizes instances close to the decision boundary.

Oversampling techniques will instead increase the instance count of the minority (positive) class to balance it with the majority class. Random oversampling replicates positive class instances to match the majority class, which risks over-fitting the classifier. To avoid this effect, the Synthetic Minority Over-Sampling Technique (SMOTE) [26] creates new synthetic instances by interpolating close minority class instances.

Finally, *ensemble learning approaches* [1] are usually performed in conjunction with one of the three other approaches to optimize the combination of classifiers. The idea behind ensemble learning [27,28] is that classifiers can be combined to improve overall performance. The key to effective ensemble learning is to combine a set of diverse classifiers [27], and the most common techniques to generate an ensemble are Bagging [29], Boosting [30] and random subspaces [31]. The literature details various ensemble learning approaches for imbalanced data. Examples of cost sensitive ensemble learning are AdaCost [32] and the AdaC family [33]. One Boosting based approach is the SMOTEBoost [34] algorithm, which adds synthetic instances of the minority class, associating different weights to original and synthetic instances, which are refined over several iterations when new synthetic samples are added.



(a) Unknown underlying distribution. (b) Data distribution of samples. (c) Approximated class models.

Fig. 1. (a) Illustration of the underlying distribution for a two class problem in a 1-D feature space. (b) A histogram representation of the data distribution obtained by discretely sampling this underlying distribution (equiprobable, 100 data samples per class) for classier design. (c) The Gaussian model estimated from data samples for classification. The balanced data used for classifier design provides limited information to estimate classifier parameters that generalize well during operations.

Bagging based ensemble approaches are popular, owing to the ease of integration of data sampling techniques and to good generalization power achieved [1]. SMOTEBagging [35] uses SMOTE resampling of the minority class with varying rate from the first to the last iteration, while majority class instances are bootstrapped. Under-sampling techniques may also be used to perform Bagging, sampling instances from the majority class and, optionally, sampling instances of the minority class for enhanced diversity. The *Asymmetric Bagging* algorithm [36] falls in this category, as well as the approach used in [37] to create ensembles for incremental learning of imbalanced data.

While the above techniques have been used to design monolithic classifiers or ensembles for problems with class imbalance, they make the ideal assumption that class imbalance of training data is representative of class imbalance associated to actual operational data. Therefore, optimizing classification systems under this assumption is likely to provide sub-optimal performance. Adapting the classification system during operations (to the actual class imbalance) should improve the system performance. In many applications (like face re-identification) recognition is performed with limited positive (or target class) samples. This paper focuses on adapting the selection and fusion of classifiers ensembles according to the class imbalance observed during operations.

2.2. Evaluation of binary classifiers

The performance of binary classifiers is commonly evaluated using the *Receiver Operator Characteristics* (ROC) [11] analysis, which is based on two intra-class measures, the true positive rate $tpr = TP/(TP + FN)$ (proportion of correct positive class predictions) and the false positive rate $fpr = FP/(FP + TN)$ (proportion of incorrect negative class predictions), derived from the confusion matrix in Table 1. ROC graphs display the entire range of *tpr* and *fpr* values to obtain different operational points and avoid committing to specific decision thresholds and classification costs. Given a classifier's evaluation of a data set, each (tpr, fpr) pair in a ROC graph represents a different decision threshold for one soft classifier, an operational point, and the empirical ROC curve is obtained by connecting the observed pairs in the graph:

Assume two operational points, op_1 and op_2, each a classifier or an ensemble of classifiers. It is said that op_1 is superior to op_2 ($op_1 \succ op_2$) if $fpr_a < fpr_b$ and $tpr_a > tpr_b$. Otherwise, if neither op_1 or op_2 are superior and they are not inferior to any other operational points, these points represent different (tpr, fpr) tradeoffs and belong to the ROC convex hull. If a ROC curve has $tpr > fpr$ for all its points, then it is a proper ROC curve. An empirical ROC curve approaches the true ROC curve as the number of (tpr, fpr) pairs on it approaches infinity.

Two different classifiers may be compared in the ROC space using the scalar measure known as the *area under the curve* (AUC). It measures classifier performance over the entire range of *fpr* values. The classifier with the higher AUC is expected to perform better. Both *tpr* and *fpr* are intra-class measures, so ROC curves are insensitive to class imbalance changes and misclassification costs. However, some real world problems require the commitment to specific scenarios.

Table 1
Confusion matrix for a binary classifier.

	Actual positive	Actual negative
Predicted positive	True Positive (TP) count Correct positive predictions	False Positive (FP) count Incorrect positive predictions
Predicted negative	False Negative (FN) count Incorrect negative predictions	True Negative (TN) count Correct negative predictions

Given an imbalance in class distributions, the *Precision-Recall Operating Characteristic* (PROC) space [12] (also known as the P-R space) focuses on an inter-class measure, classifier $precision = TP/(TP + FP)$ (proportion of correct positive predictions against the total positive predictions), which is related to classification accuracy, as well as *recall* (the same as *tpr*). PROC graphs represent classifier performance regarding data imbalance through precision, the proportion of correct positive predictions. Therefore, the same classifier has different PROC curves when evaluated using data with different class imbalance. In contrast, the ROC curves would be equivalent as both *tpr* and *fpr* are insensitive to class imbalance. Davis and Goadrich discussed [13] the equivalence between dominating operational points in the ROC and PROC spaces, from which they derived a methodology to find the PROC achievable curve (the analogous to the ROC convex hull).

The F-measure (Eq. (1)) is a scalar performance measure is commonly used to characterize the PROC space. It grows when both *precision* and *recall* increase simultaneously. When $\beta = 1$, the F_1 score is the harmonic mean of precision and recall. It is often used on information retrieval, which faces severe class imbalances, and is based on Rijsbergen's effectiveness measure [38].

$$F_\beta = (1 + \beta^2) \times \frac{precision \times recall}{(precision \times \beta^2) + recall} \tag{1}$$

2.3. Selection of decision thresholds

Fig. 2a shows a synthetic example that is used throughout the paper. The bi-dimensional data (shown here only in the first dimension) is generated from two spherical Gaussian distributions. $\mathcal{N}_{+(\boldsymbol{\mu}_+, \mathbf{I})}$ indicates the positive (minority) class, while $\mathcal{N}_{-(\boldsymbol{\mu}_-, \mathbf{I})}$ indicates the negative (majority) class, where $\boldsymbol{\mu}_+ = (0, 0)$ and $\boldsymbol{\mu}_- = (1.5, 1.5)$. These distributions are used to generate data sets with the different class proportions shown in Table 2. Balanced training data (`train`) 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. Thus, c_1 and c_2 are linear decision boundaries perpendicular to the x and y axis, respectively. The three other data sets (`val`, `opt` and `test`) are used with different class imbalance levels through random subsampling, to provide variable levels of class imbalance that range from 1/1 to 1/1000. The `opt` data is used in Section 3, while the validation data `val` is used as independent data to select an operational point, either a single classifier decision threshold or a classifier Boolean combination [39,40]. Finally, the `test` data is used to evaluate classification performance.

Given the soft binary classifier c_1, a decision threshold γ is selected to define an operational point that provides a performance trade-off regarding positive and negative classes. An operational threshold γ is often selected with independent validation data (`val`) once the classifier has been designed. The optimal decision boundary for a classifier is selected to minimize the probability of error according to the Bayes theory, which is equivalent to the equal error rate (EER) when the positive and negative classes are balanced. For c_1, the optimal decision boundaries for both $\lambda = 1/1$ and $\lambda = 1/10$ ($\pi_p = 200$ and $\pi_n = 2000$) correspond to the decision thresholds γ_1 and γ_2 in Fig. 2.

Assume that class imbalance is known. Evaluating the performance of classifier c_1 produces the ROC and PROC plots in Fig. 2b and c, respectively. Although the ROC curve is insensitive to changes in class proportions, PROC curves demonstrate the impact of class imbalance on classification performance. Moreover,

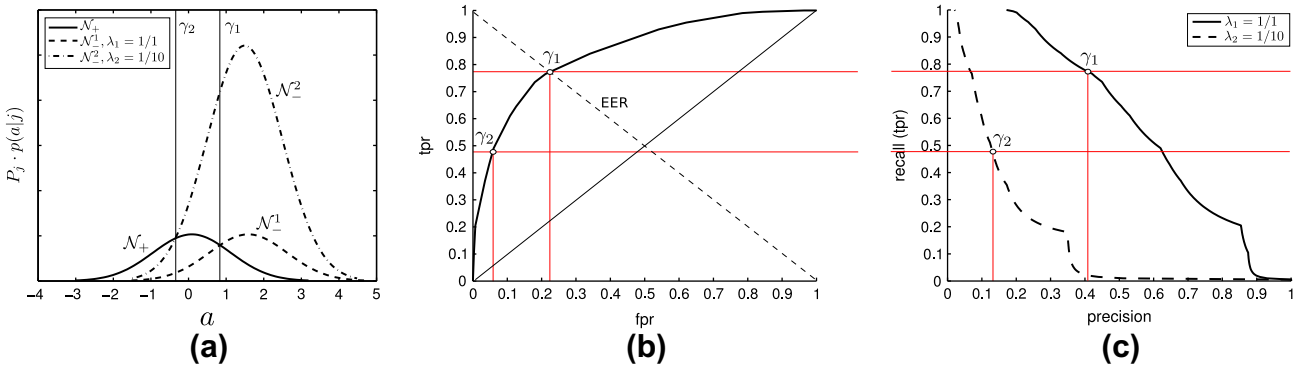


Fig. 2. A synthetic example – (a) distribution of `val` data in bi-dimensional space as perceived in the abscissa axis, for both $\lambda_1 = 1/1$ and $\lambda_2 = 1/10$ class imbalance levels. The optimal decision boundaries for λ_1 and λ_2 are decision thresholds γ_1 and γ_2, respectively. Projection of thresholds onto the ROC (b) and inverted PROC (c) graphs obtained for classifier c_1. Selecting the optimal decision boundary results in large variations of the observed *tpr* and *fpr* values.

Table 2
Proportions of the negative and positive samples generated for data sets used for synthetic proof of concept experiments. Random sub-sampling of negative class samples is used to generate imbalanced data (maximum of $\lambda = 1/1000$) for the `val`, `opt` and `test` data sets.

Dataset	Positive samples – π_p	Negative samples – π_n	Skew: λ
`train`	100	100	1/1
`opt`	200	200,000	1/1 to 1/1000
`val`	200	200,000	1/1 to 1/1000
`test`	1000	1,000,000	1/1 to 1/1000

selecting the optimal decision boundary provides a large variation in the *fpr*, *tpr* and precision measures.

For example, in video surveillance applications, an acceptable *fpr* is set by the human operator, projecting it to an operational point. Thus, the decision threshold is a variable defined by the target *fpr* and class imbalance. Using data in Table 2, a decision threshold set Γ for c_1 at $fpr = 5\%$ is selected for each class imbalance level ($\lambda_1 = 1/1$ and $\lambda_2 = 1/10$), such as that $\Gamma = \{\gamma'_1, \gamma'_2\}$. Regardless of class imbalance, c_1 provides the same *fpr*, even though the PROC space indicates decreasing levels of performance as class imbalance increases. Fig. 3a and b details the operational points in the ROC and PROC spaces, which allows a consistent behavior for a video surveillance application. Fig. 3c complements the example by projecting the decision thresholds back onto the underlying distribution space.

Evaluating classification performance on the previously unseen `test` data with λ varied from 1/1 to 1/1000 provides the results in Table 3. Selecting a fixed threshold γ with balanced `val` data (λ_1) provides actual (test set) *fpr* values that tend to increase beyond 5% with higher class imbalance levels. On other hand, selecting a decision threshold according to a target imbalance level, using `val` data matching the testing data ($\lambda \in \{1/10, 1/100, 1/1000\}$), provides *fpr* values that are closer to the target $fpr = 5\%$, while improving precision and F_1 values.

Results suggest that given a target *fpr*, the automatic selection of a threshold γ is better by using imbalanced validation data that corresponds to the skew of operational data. In this situation, actual (operational) *fpr* values should be closer to target *fpr* values, while precision and F_1 measures are higher. Accounting for class imbalance and the PROC space can lead to better classifier ensembles where *fpr* values are closer to the target values.

2.4. Estimation of class imbalance

In literature, some approaches have been proposed to estimate class imbalance, and these are useful to adapt classifier ensembles. Using operational data to re-estimate classifier parameters is referred to as transductive learning. A direct approach is to use the classifier outputs to estimate class imbalance. This approach relies on the classifier performance, which may be affected by imbalance. Therefore it does not necessarily provides reliable estimations. Yang and Zhou [41] proposed the Online Expectation–Maximization (OEM) algorithm for transductive learning, using classifier posterior probabilities in conjunction with the EM algorithm to iteratively estimate class priors and adapt classifier parameters.

A transfer estimation of class priors is discussed in [42], it extends the OEM algorithm into the Transfer Online Expectation–Maximization (Tr-OEM) algorithm. Transfer learning is a technique

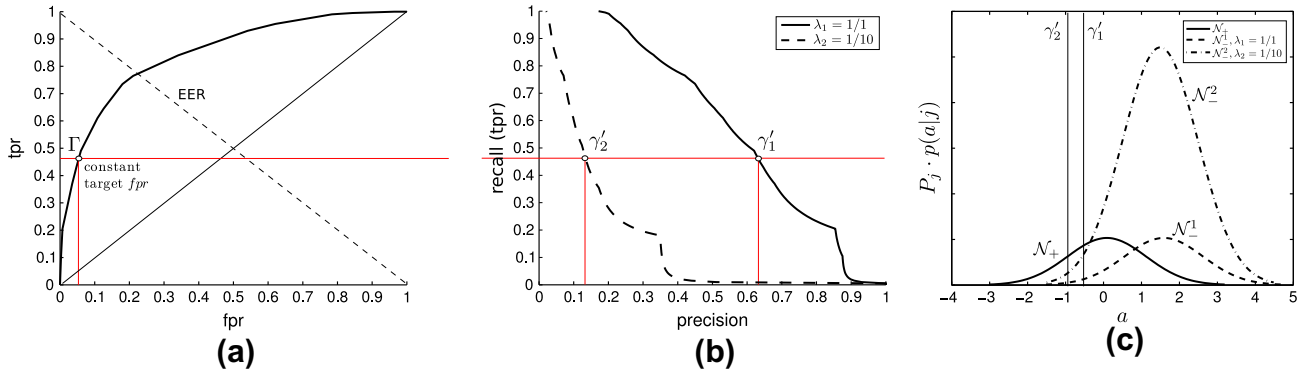


Fig. 3. ROC (a) and inverted PROC (b) graphs comparing decision thresholds obtained with `val` data on a fixed $fpr = 5\%$ at two different levels of class imbalance (1/1 and 1/10). Decision thresholds $\bar{\Gamma} = \{\gamma'_1, \gamma''_2\}$ (c) as perceived by c_1 in the abscissa axis, for both $\lambda_1 = 1/1$ and $\lambda_2 = 1/10$ class imbalance levels.

Table 3
Performance on unknown test data for decision thresholds γ at $fpr = 5\%$, using both balanced validation data ($\lambda = 1/1$) and using imbalanced data that matches the testing data class imbalance (from 1/1 to 1/1000). The left half of the table represents the selection of a single decision threshold γ, with balanced `val` data. Each row in the right half uses a decision threshold obtained with `val` data matching the class imbalance indicated in the column γ.

Test set λ	Balanced validation ($\lambda = 1/1$)					Imbalanced validation (λ varies from 1/1 to 1/1000)				
	γ	*fpr* (%)	Recall (%)	Precision (%)	F_1	γ	*fpr* (%)	Recall (%)	Precision (%)	F_1
1/1	0.802	4.90	45.20	90.22	0.602	0.802	4.90	45.20	90.22	0.602
1/10	0.802	5.67	47.50	45.59	0.465	0.811	5.26	46.30	46.81	0.466
1/100	0.802	5.82	46.70	7.43	0.128	0.817	5.14	44.50	7.97	0.135
1/1000	0.802	5.69	46.90	0.82	0.016	0.818	4.99	44.60	0.89	0.018

from data mining, which uses data from related domains to improve learning when few samples from the target domain are available. They propose using part of the training data, with known labels, to estimate the class priors. For one unknown sample, the approach uses the true label of its known neighbors to calculate the class frequencies, which are used to approximate class priors.

For classification problems, two data sets have the same class imbalance when their histograms in the feature space are comparable (i.e., their data distribution is comparable). The approach described in [14] uses the Hellinger distance in the feature space to select, from several labeled data sets with known class imbalances, which has the closest class imbalance to unlabeled operational data.

Assume a set of unlabeled operational data (`uod`) and several labeled reference data sets (`ld`), each one characterized by a different level of imbalance. The Hellinger distance $H(\texttt{ld}, \texttt{uod})$ indicates the dissimilarity between data distributions of those two sets using histogram representations of the feature space. The labeled set with the smallest Hellinger distance to *uod* is the one with the closest level of class imbalance. For a given number of features f in data patterns and bins b in the histogram, the Hellinger distance is calculated as:

$$H(\texttt{ld}, \texttt{opd}) = \frac{1}{f}\sum_{j=1}^{features}\sqrt{\sum_{i=1}^{bins}\left(\sqrt{\frac{|\texttt{ld}_{j,i}|}{|\texttt{ld}|}} - \sqrt{\frac{|\texttt{uod}_{j,i}|}{|\texttt{uod}|}}\right)^2} \quad (2)$$

BC techniques combine binary decisions of classifiers, and the resulting Boolean fusion function and thresholds, provide a binary decision. Approaches that rely on posterior probabilities of the classifiers are not suitable to estimate class imbalance. Therefore, using a density-based approach in the feature space (like measure Hellinger distance between data sets), is appropriate to estimate class imbalance for multi-classifier FR systems based on BC.

3. Boolean combination of classifiers

Boolean combination (BC) are versatile techniques for threshold-optimized fusion of crisp and soft 1- or 2-class classifiers at the decision level [43,44]. A soft classifier c_i produces a binary decision when its normalized output score is compared to a threshold $0 \leqslant \gamma_i \leqslant 1$. This decision c_{i,γ_i}, provides a performance trade off between positive and negative classes (for instance, a point in the ROC space). Given a set of decision thresholds Γ, the BC of two soft classifiers c_i and c_j is the fusion of all c_{i,γ_i} and c_{j,γ_j} using Boolean operations. Given a classifier c, its ROCCH is the ROC curve composed of the vertices c that maximize the area under the ROC curve (AUC). Each EoC vertice is based on decision thresholds for classifiers and a Boolean function, and corresponds to an EoC. Selecting the superior operational points in the decision space (for instance, the ROC convex hull or the PROC achievable curve) defines the best performance trade off.

The rest of this section reviews algorithms for BC of soft classifiers in the ROC space, and presents a methodology to select an operational point for a desired *fpr*. The relevance of specialized BC techniques for imbalanced data is then illustrated empirically.

3.1. Boolean combination algorithms

The Boolean conjunction (AND) and disjunction (OR) fusion functions were first used in BC techniques to combine crisp detectors that are conditionally-independent in [45]. Tao and Veldhuis proposed a threshold-optimized decision-level fusion technique based on AND and OR functions that is optimal in the Neyman–Pearson sense [44].

Haker et al. [46] used a set of Boolean operators with the conjunction ($\wedge$) and disjunction ($\vee$) to combine ROC space vertices of two soft classifiers. This BC technique consists in defining a set of thresholds that are applied to the output scores of two classifiers c_i and c_j, to obtain combined operational points. The drawback to this approach is that it generates all possible combinations to select the superior operational points – vertices in the ROCCH. By extending the algorithm for more classifiers, the number of combinations explored grows exponentially, as well as memory usage and processing time. Finally, Haker's approach assume that classifiers c_i and c_j are conditionally independent, which is not necessarily true for real-world problems. Khreich et al. [10] proposed the Iterative Boolean Combination (IBC) which uses 10 Boolean functions to combine classifiers in the ROC space. IBC is computationally efficient for several classifiers – instead of generating the entire set of combination points, it incrementally combines classifiers one by one, pruning points in the decision space by selecting only points belonging to the ROCCH each time. Also, the set of Boolean operations used by IBC does not assume that classifiers are conditionally independent.

Section 2.3 has shown the advantage of using data to select a classifier decision threshold that follows the same level of imbalance as seen during operations. BC of classifiers in the decision space should therefore be performed using imbalanced data that corresponds to operational data, allowing to generate better operational points and performing selection and fusion of the most suitable ensembles.

3.2. Selection of operational point

The outcome of a BC technique is a set EoCs, each EoC one corresponding to a vertice of the ROCCH. After performing BC, the next step is to define an operational point for the specific application. A general approach to select an operational point is to choose the EoC in the BC that provides the best trade off between *tpr* and *fpr* values, but for a specific application, the operational point is typically selected for a desired *fpr* value, using validation data. However, it is unlikely that an EoC produced by BC will correspond to the desired *fpr* value.

Scott et al. in [47] proposed a method to interpolate between two consecutive vertices (EoCs) in the ROCCH, E_i and E_j, to realize an operational point E_k between the two original EoCs (Fig. 4). To classify input samples, the interpolation method alternates

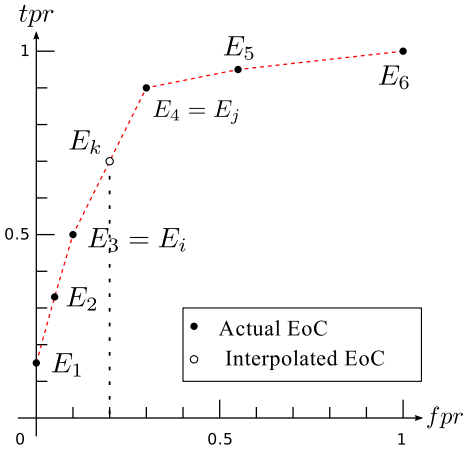


Fig. 4. Illustration of the ROCCH produced for a BC with 6 vertices, $BC = \{E_1, E_2, \ldots, E_6\}$. The operational point (ensemble) E_k for the desired $fpr = 20\%$ lies between two vertices, E_i (E_3) and E_k (E_4), and is obtained by interpolation.

between the decisions provided by E_i and E_j for each sample. The probability of selecting one of the two vertices is determined by the distance of E_k to the vertexes E_i and E_j:

$$P(E_k = E_j) = \frac{fpr_{E_k} - fpr_{E_j}}{fpr_{E_j} - fpr_{E_i}} \tag{3}$$

$$P(E_k = E_i) = 1 - P(E_k = E_j) \tag{4}$$

assuming that fpr_{E_k} is the target fpr value.

3.3. Boolean combination with imbalanced data

Section 2.3 describes the selection of a decision threshold for a single classifier under different levels of class imbalance using the PROC space. This section extends this concept for the selection of an operational point (an EoC) from BCs produced for different levels of class imbalance in the set $\Lambda = \{1/1, 1/10, 1/100, 1/1000\}$. Classifiers c_1 and c_2 (see Section 2.3) are combined with IBC [10], using optimization `opt` and an operational point is selected with validation `val` data (Table 2) for each level of class imbalance in Λ.

The BC optimization data `opt` is used to define the decision thresholds in Γ (with $|\Gamma| = 100$, a parameter for IBC) that are used to produce a set of BCs E, where each BC is in turn a set of EoCs (vertices in the ROC/PROC space). Next, operational points $E_{k,\lambda}$ are selected for each BC in E, using validation data the different class imbalance levels. The ROC and PROC curves for different levels of class imbalance are shown in Fig. 5, with operational points selected for a target $fpr = 5\%$. As with the selection of a decision threshold for a single classifier, optimizing the BC and selecting an operational point using imbalanced data provides a constant fpr, regardless the different levels of performance indicated in the PROC space.

Table 4
Selected operational points for a target $fpr = 5\%$ and different class imbalance levels λ to perform BCs. EoCs $E_{i,\lambda}$ and $E_{k,\lambda}$ are consecutive vertexes in the ROC space, and are represented by the BC of the classifiers and their decision thresholds. The selected operational point E_k is obtained with the method described in Section 3.2, according to probabilities $P(E_{k,\lambda} = E_{i,\lambda})$ and $P(E_{k,\lambda} = E_{j,\lambda})$.

Skew λ	$E_{i,\lambda}$	$E_{j,\lambda}$	$P(E_{k,\lambda} = E_{i,\lambda})$ (%)	$P(E_{k,\lambda} = E_{j,\lambda})$ (%)
1/1	$c_{1,0.457} \wedge c_{2,0.702}$	$c_{1,0.536} \wedge c_{2,0.261}$	33.33	66.67
1/10	$c_{1,0.806} \vee c_{2,0.903}$	$c_{1,0.806} \vee c_{2,0.748}$	85.32	14.68
1/100	$c_{1,0.495} \wedge c_{2,0.527}$	$c_{1,0.495} \wedge c_{2,0.431}$	2.70	97.30
1/1000	$c_{1,0.388} \wedge c_{2,0.668}$	$c_{1,0.388} \wedge c_{2,0.515}$	36.29	63.71

Performance is evaluated on the `test` data set in two stages. First, a single BC is performed with balanced ($\lambda^{bal} = 1/1$) `opt` data, and an operational point at $fpr = 5\%$ is selected with balanced `val` data. This BC is evaluated with `test` data at different levels of imbalance in $\Lambda = \{1/1, 1/10, 1/100, 1/1000\}$. In the second stage, the same `test` data is evaluated with BCs performed with `opt` data that is imbalanced following Λ, and operational points are selected for $fpr = 5\%$ using `val` data with the class imbalance levels matching those used in `test`. Resulting operational points for each class imbalance level in Table 4 are obtained using the method described in Section 3.2. These results confirm trends observed in Section 2.3, that the most suitable operational point changes for different levels of class imbalance for a fixed $fpr = 5\%$.

EoC performance on the `test` data set is shown in Table 5 for each class imbalance level $\lambda \in \Lambda$. The left side of the table represents performance with a single operational point, produced with balanced BC optimization data (`opt` and `val`), which refers to the first line ($\lambda = 1/1$) in Table 4. The right side of this table presents results using imbalanced BC `val` optimization data which follows the same class imbalance as the `test` data. Operational points produced with imbalanced data are closer to the targeted

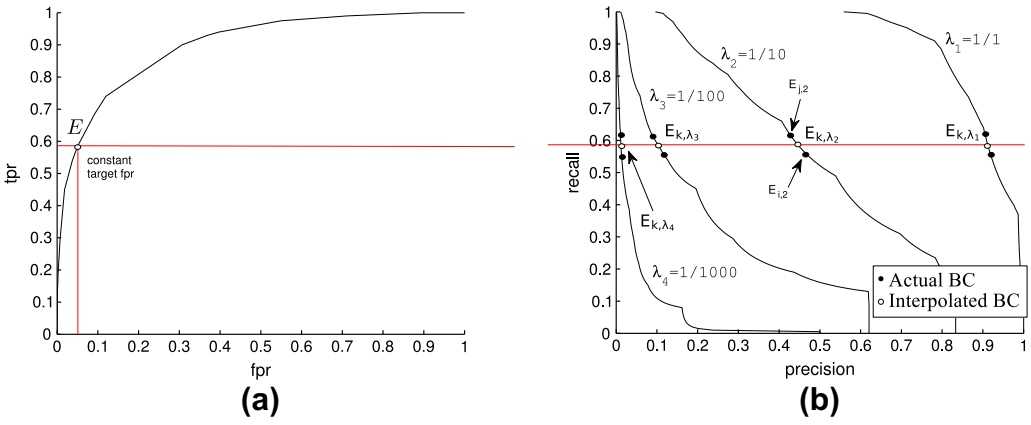


Fig. 5. ROC and inverted PROC graphs obtained with four BC optimization data `opt` sets, corresponding to four different levels of class imbalance $\lambda_1 = 1/1, \lambda_2 = 1/10, \lambda_3 = 1/100$ and $\lambda_4 = 1/1000$, for a target $fpr = 5\%$. ROC curves in 5a are overlapping, while PROC curves in 5 are different for each level of class imbalance. EoCs are selected using the method described in Section 3.2 for each class imbalance level, where EoC_j and EoC_i are the closest adjacent vertices in the ROC space, and EoC_k is the resulting EoC for operation.

Table 5
Average performance over 10 replications of BCs at $fpr = 5\%$ on various levels of test data class imbalance. Standard deviation values are indicated between parenthesis.

Skew λ	Balanced validation				Imbalanced validation			
	fpr	Recall (*tpr*)	Precision	F_1	*fpr*	Recall (*tpr*)	Precision	F_1
1/1	5.78% (0.011)	58.82% (0.036)	91.08% (0.014)	0.714 (0.026)	5.78% (0.011)	58.82% (0.036)	91.08% (0.014)	0.714 (0.026)
1/10	5.78% (0.008)	58.61% (0.027)	50.52% (0.025)	0.542 (0.015)	5.09% (0.004)	58.59% (0.022)	53.53% (0.022)	0.559 (0.020)
1/100	5.88% (0.007)	58.65% (0.034)	9.13% (0.007)	0.158 (0.010)	5.11% (0.001)	59.25% (0.012)	10.40% (0.003)	0.177 (0.004)
1/1000	5.85% (0.007)	57.76% (0.034)	0.99% (0.001)	0.019 (0.001)	5.01% (0.000)	57.92% (0.031)	1.14% (0.001)	0.022 (0.001)

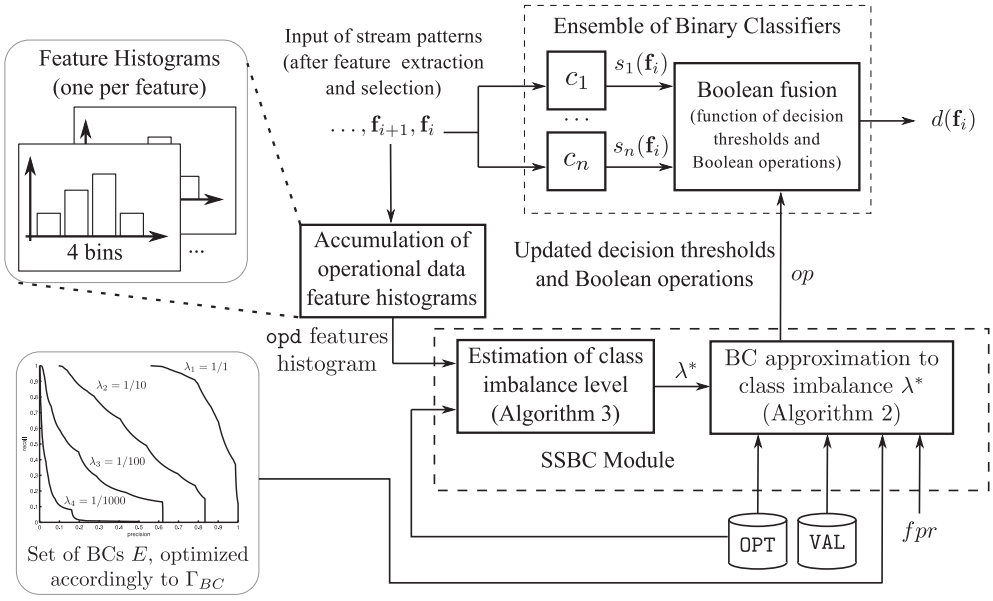


Fig. 6. Architecture that adapts a BC of classifiers to imbalanced class distributions.

$fpr = 5\%$ on the `test` data set, and the scalar precision and F_1 values are consistently higher.

Using imbalanced `val` data for BC of classifiers, where class proportions correspond to operational (test) data, provides a better choice of operational point and a higher level of performance for a fixed *fpr* value. These results support the idea of adapting BC (the selection and fusion of classifiers) to changing class proportions, although the computational cost of re-calculating the BC for a specific level of imbalance is high. To alleviate this problem, the next section proposes a computational efficient approach to estimated class proportions during operations, and to approximate BC from an initial set of pre-calculated BCs over several pre-defined class imbalance levels.

4. Adaptive skew-sensitive boolean combination

In this section, an adaptive Skew-Sensitive Boolean Combination (SSBC) technique is proposed for the selection and fusion of the most accurate ensembles according to the class imbalance observed from input data during operations. It is assumed that samples of operational data closest to class imbalance level λ^*, from a set of known class imbalance levels Λ. The class imbalance level values in Λ range from $\lambda^{bal} = 1/1$ to a maximum level λ^{max}. These are the class imbalance levels observed, for instance, by a face recognition system. Specific class imbalance levels in Λ, between λ^{bal} and λ^{max}, can be obtained from a single data set with λ^{max} class imbalance, through random under sampling of the majority class. To reduce memory consumption and computational complexity to generate BCs, a smaller subset of class imbalance levels $\Lambda_{BC} \subset \Lambda$ is initially defined to generate a set E of BCs, instead of storing a BC for every class imbalance level in Λ. This set Λ_{BC} contains evenly distributed class imbalance levels from Λ, between λ^{bal} and λ^{max}.

If the estimate $\lambda^* \in \Lambda$ matches a class imbalance level in Λ_{BC} (the set of pre-optimized BCs E), then the EoC $E_{\lambda^*} \in E$ is directly selected for operations. However, when λ^* is not in Λ_{BC} (i.e., no BC is pre-optimized for λ^*), there are two class imbalance levels λ^i and λ_j in Λ_{BC} that are adjacent to λ^* ($\lambda^i < \lambda^* < \lambda^j$).

Both BCs E_{λ^i} and E_{λ^j} have vertices in the PROC space that are close to the unknown E_{λ^*}. In this situation, an approximation can be obtained from the combination of vertices in E_{λ^i} and E_{λ^j}. Once combined, the vertices (EoCs) have their performance (*tpr* and *fpr*) evaluated with imbalanced reference data matching λ^* and the superior points (the ROCCH) are selected as E_* for operation.

4.1. Proposed architecture

Fig. 6 presents the block diagram of an adaptive classification system based on the new skew-sensitive BC (SSBC) technique described above. It allows for adaptive selection ensembles of binary classifiers, based on its estimation of class imbalance. Assume for example a stream of facial regions of interest captured within video frames by some FR system. Facial regions are transformed by feature extraction and selection to patterns for classification, and

also accumulated over time into feature histograms (opd) used to estimate λ^*, the closest level in a set Λ of known class imbalance levels. Since this estimate may change over time, and BC is a computationally intensive task, the SSBC technique is proposed to cost-effectively adapt ensembles of classifiers. BCs are approximated from adjacent levels of class imbalance λ^i and λ^j using validation data following λ^*, the level of class imbalance estimated from the operational data. The approach can approximate the BC up to a maximum λ^{max} class imbalance level.

The approach uses a set of known levels of class imbalance, $\Lambda = \{\lambda^{bal} = 1/1, \ldots, \lambda^{max}\}$, to which the system can adapt, and a subset $\Lambda_{BC} \subset \Lambda$ that is selected to optimize an initial set of BCs E. The set Λ_{BC} contains evenly distributed intermediate class imbalance levels between λ^{bal} and λ^{max} inclusively. The SSBC approach uses OPT and VAL: data sets that follow the levels of class imbalance in Λ. Data sets in OPT and VAL are generated from imbalanced reference data that follows λ^{max} through random under sampling. The target minority class is held fixed, while those from the non-target class are grown through random sub-sampling. This process generates different data sets with levels of class imbalance from $\lambda^{bal} = 1/1$ to λ^{max}, as defined in Λ.

Once a pool of binary classifiers $C = \{c_1, \ldots, c_n\}$ is generated using balanced data, ensemble selection and fusion is performed using the Iterative Boolean Combination (IBC) techniques [10]. BC of pool C is performed during the design phase using the levels of imbalance in Λ_{BC} (see Algorithm 2). Each BC in E is optimized for corresponding class imbalance level in Λ_{BC}, with data matching the same class imbalance in OPT and using t decision threshold. Once BC is performed, the approach assumes that data is balanced and operates at $\lambda^{init} = \lambda^{bal} = 1/1$. An operational point *op* for a target *fpr* from $E_{\lambda^{init}}$ is selected using the data set $val_{\lambda^{init}}$.

During system operation, the feature histogram of operational data is accumulated over time to periodically estimate the closest level of class imbalance $\lambda^* \in \Lambda$ (see Algorithm 1). The opd feature histogram is compared to data sets in OPT (each following a class imbalance defined in Λ) using the Hellinger distance. The class imbalance level associated to the data set in OPT with lowest Hellinger distance to opd corresponds to λ^*, the closest class imbalance estimate in OPT.

Once the closest class imbalance level λ^* is estimated from the levels available in Λ, the BC is approximated using Algorithm 3. This algorithm takes the class imbalance level λ^* and determines if there is a BC already available in E, or if it must be approximated from two adjacent BCs in E. The approximation combines the vertices (EOCs) of two BCs, optimized for adjacent class imbalance levels, and select the superior points (the ROCCH) using optimization data $\mathtt{opt}_{\lambda^{init}}$. Once the BC E_{λ^*} is selected (directly or through optimization), an updated operational point *op* is selected with $\mathtt{val}_{\lambda^{init}} \in$ VAL and used to update the Boolean fusion function that provides the final decision.

Assuming an ensemble-based classification system that already uses a pool of diverse classifiers C, the overhead of the SSBC in memory is as follows. Reference data (OPT and VAL) required to adapt the system is directly related to the cardinality of Λ. The larger $|\Lambda|$, the more memory is required to store reference data for each class imbalance level. Similarly, the cardinality of Λ_{BC}, defines the number of BCs in E (decision thresholds and Boolean functions between classifiers). Finally, the accumulated histogram of operational data size is defined by the number of features extracted and the bins required for the histogram. The most important memory footprint is related to storage of reference data and designing Λ. However, the accuracy of approximations is directly related to the number of pre-optimized BCs.

The remainder of this section details the proposed method to approximate BC to new class imbalance levels, followed by the approach to estimate the closest class imbalance level in Λ from unlabeled operational data. Finally, a proof of concept experiment is performed to illustrate the potential effectiveness of SSBC in adapting BC to changes in class imbalance.

4.2. Estimation of the closest class imbalance level λ^*

Section 2.4 introduced the Hellinger distance H in Eq. (ref-eq:hvalopd) which was used by [14] to estimate the closest class imbalance level based on known imbalanced data. Assume the set Λ of class imbalance levels and the set of data sets OPT, where each data set in OPT follows one different class imbalance level in Λ. Algorithm 1 details the process to estimate λ^*, the class imbalance level in Λ which has the closest class proportions to unlabeled operational data opd.

The algorithm first calculates the Hellinger distance between the features histogram opd and all data sets in OPT (each data set following a class imbalance level in Λ). The labeled data set in OPT with the smallest Hellinger distance to opd indicates λ^*, the class imbalance level in Λ which is closest to the unknown class imbalance level associated to operational data. Given L_+ the positive class samples in the reference data OPT (fixed, regardless of class imbalance), the bin number b used to calculate the Hellinger distance is $b = \lfloor \sqrt{L_+} \rfloor$.

Algorithm 1. Estimation of the closest level of class imbalance λ^* from a set of data sets OPTand unlabeled operational data opd.

Data: Data set OPT, operational data features histogram opd and b bins
Result: Estimation of the closest class imbalance level λ^* from OPT
$min = \infty$;
$\lambda^* = 0$;
for allopt $\in$ **OPT do**
  $hd = H(\mathtt{opt}, \mathbf{opd}, b)$;
  If $hd < min$
    $min = hd$;
    Set λ^* to class imbalance level of **opt**;
end

4.3. Approximating BCs to new class imbalances

Given the known set of class imbalance levels Λ, the set of BCs E created with Algorithm 2, and assuming λ^*, the level of class imbalance estimated from operational data of class imbalance levels in Λ. The procedure to approximate BCs to the class imbalance level λ^* is indicated in Algorithm 3 and graphically represented in Fig. 7, where the approximated BC is indicated by the dashed line in Fig. 7b.

Algorithm 2. Initial BC design for the SSBC technique described in Algorithm 3.

Data: Pool of classifiers C, maximum number of decisions thresholds t, data sets OPT and VAL and the target *fpr*
Result: Set of BCs E and the initial operational point *op* (an EoC) for $\lambda^{init} = 1/1$.
$E = \emptyset$;
forall opt $\in$ OPT with a level of class imbalance in Λ_{BC}**do**
  $E = E \cup \{IBC(C, t, \mathtt{opt})\}$;
end
Select $E_{\lambda^{init}} \in E$ for $\lambda^{init} = 1/1$;
Select $op \in E_{\lambda^{init}}$ for the target *fpr* with $val_{init} \in$ VAL;

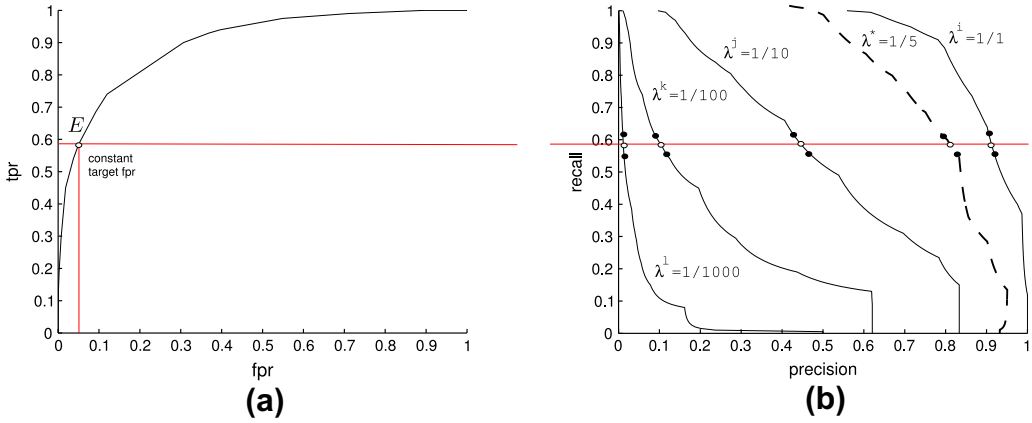


Fig. 7. ROC and inverted PROC graphs obtained with the validation data `val` when fixing the target *fpr* = 5%. Given the BCs optimized for the levels of class imbalance in Λ (solid lines), the SSBC approach can approximate the BC for $\lambda^* = 1/5$ (dashed line) from the EoCs in the adjacent BCs $\lambda^i = 1/1$ and $\lambda^j = 1/10$.

When $\lambda^* \in \Lambda_{BC}$, the BC E^* is selected directly from E. Otherwise, the BC is estimated as follows. First, the adjacent class imbalance levels $\lambda^i, \lambda^j \in \Lambda_{BC}$ are determined. Next, the `opt`* data set is selected from `OPT`, following the same class imbalance level as λ^*. Then EoCs (vertices in the ROCCH) in both E_{λ^i} and E_{λ^j} are combined, and only the points projected in the ROCCH with the `opt`* data set are kept in E^*. Finally, an operational point is selected for the target *fpr* using `val`*, validation data that in `VAL` follows the level of class imbalance λ^*.

Algorithm 3. SSBC technique for adapting BC for a new λ^* class imbalance level.

Data: set of BCs E, set of class imbalance levels Λ_{BC}, data sets `OPT` and `VAL`, the estimated class imbalance $\lambda^* \in \Lambda$ and the target *fpr*.
Result: Operational point *op* for the target *fpr*.
$E^* = \emptyset$
If $\lambda^* \in \Lambda_{BC}$ **then**
  $E^* = E_{\lambda^*}$;
end
else
  Select $\lambda^i, \lambda^j \in \Lambda_{BC}$, such as that $\lambda^i < \lambda^* < \lambda^j$;
  Select `opt` $^* \in$ `OPT`, following λ^*;
  $E^* = ROCCH(E_{\lambda^i} \cup E_{\lambda^j}, \texttt{opt}^*)$;
end
Select `val` $^* \in$ `VAL`, following λ^*;
Select $op \in E^*$ for the target *fpr* with `val`*;

To verify the proposed approach to approximate BC to new levels of class imbalance, an extension to the experiment in Section 3.3 is performed in 10 replications, with $\Lambda_{BC} = \{1/1, 1/10, 1/100, 1/1000\}$ and $\Lambda = \{1/1, 1/5, 1/10, 1/55, 1/100, 1/550, 1/1000\}$. This experiment verifies the approximation of BCs to class imbalance levels $\lambda^*_a = 1/5, \lambda^*_b = 1/55$ and $\lambda^*_c = 1/550$ (all intermediate values of levels in Λ_{BC}). For comparison purposes, the actual BC for these skew levels are also calculated.

Table 6 shows mean and standard deviation values over the 10 trials for different performance metrics. The left side of the table shows values obtained with the full BC optimization for λ^* and a target *fpr* = 5%, while the right side does the same for the SSBC technique to approximate BC to λ^*. Values obtained with both approaches are equivalent and no significant difference is observed for the selected operational point. Thus, the SSBC technique to approximate BCs to a new class imbalance λ^* can replace the costly full BC during operations.

Algorithm 3 is computationally more efficient than full BC to new class imbalance levels λ^*. For 2 classifiers and t decision thresholds, the worst case time complexity for IBC is $O(t^2)$. For the simulations in this paper ($t = 100$), about 20,000 EoC evaluations were required with IBC. The approximation strategy in Algorithm 3 requires $O(|E_{\lambda^1}| + |E_{\lambda^2}|)$ in the worst case. In simulations, there was a significant reduction to about 1% of the original computational effort. Memory requirement is also considerably smaller with Algorithm 3, requiring $O(|E_{\lambda^1}| + |E_{\lambda^2}|)$ vertices stored in memory in the worst case, against $O(t^2)$ for IBC.

4.4. Validation with synthetic data

To validate the proposed SSBC technique, a final experiment with synthetic data is performed with c_1 and c_2 (see Section 2.3). Parameters for this experiment are as follows: $\Lambda = \{1/1, 1/10, 1/50, 1/100, 1/200, \ldots, 1/1000\}$, $\Lambda_{BC} = \{1/1, 1/10, 1/100, 1/500, 1/1000\}$ and $\lambda_{init} = 1/1$. Data is divided in 6 ordered data blocks with 400 positive samples each, and class imbalance levels as follows: 1/30, 1/300, 1/600. 1/960, 1/400 and 1/120. For instance, the first block has 400 positive samples and 12,000 negative samples.

To simulate a stream of unknown samples over time, samples are presented to the system in a random order. The closest class imbalance level in Λ_{BC} is estimated in the sample interval containing the last 400 samples, at each 200 positive samples interval. The simulation is replicated 10 times to better evaluate performance and is compared to the traditional static approach of one BC optimized with balanced validation and selection data obtained through random under sampling (see Fig. 8).

Estimation of class imbalance using Algorithm 1 produces approximations over time as indicated in Fig. 9 for one random replication. The closest class imbalance level in Λ is presented as a square shaped curve, owing to the estimation in regular intervals (at each 200 positive samples). Algorithm 1 provides a proper, yet not ideal estimation of class imbalance level using the Hellinger distance, which is limited by the class imbalance levels in Λ. However, increasing the set Λ will consume more memory for E, `OPT` and `VAL`.

Considering the average performance for each data chunk in Table 7, we observe that the proposed adaptive SSBC technique outperforms the traditional static approach. The *fpr* values after

Table 6
Comparison between the proposed SSBC approach and the full BC for the target class imbalance. Values are the average performance in 10 replications of BCs at $fpr = 5\%$ on test data. Standard deviation values are indicated between parenthesis.

λ	Full BC				SSBC (approximation)			
	fpr	Recall (*tpr*)	Precision	F_1	*fpr*	Recall (*tpr*)	Precision	F_1
1/5	4.69% (0.008)	57.71% (0.030)	71.27% (0.026)	0.637 (0.016)	4.86% (0.010)	57.49% (0.031)	70.53% (0.034)	0.632 (0.018)
1/55	5.07% (0.003)	57.82% (0.019)	17.21% (0.007)	0.265 (0.008)	5.03% (0.003)	58.10% (0.019)	17.38% (0.007)	0.267 (0.008)
1/550	5.00% (0.000)	58.10% (0.018)	2.07% (0.001)	0.040 (0.001)	4.98% (0.001)	58.50% (0.013)	2.09% (0.001)	0.040 (0.001)

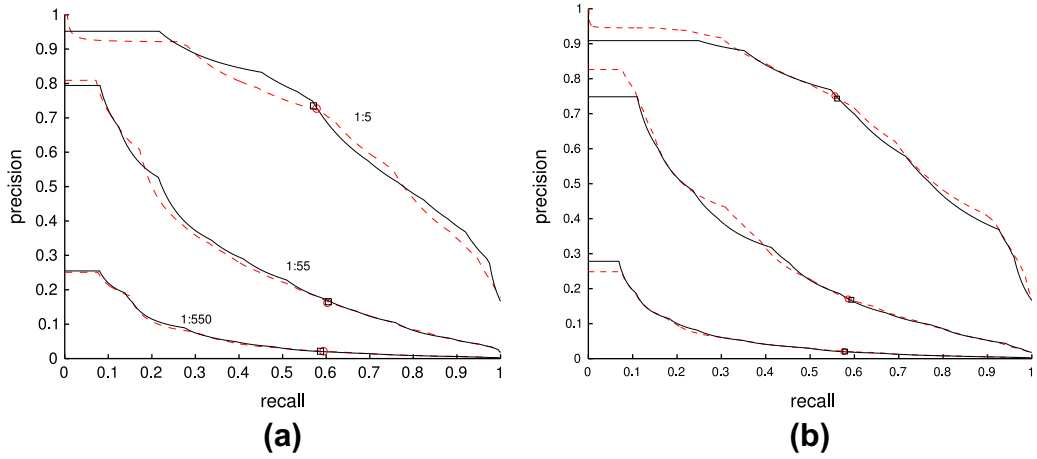


Fig. 8. BC approximated PROC curves in the `test` data set, for two different replications. Solid lines are the BC approximation and dashed lines are the full BC optimization for the target class imbalance. The circle is the selected operational point using the approximated BC, which coincides with the selected operational point with the full BC optimization (square).

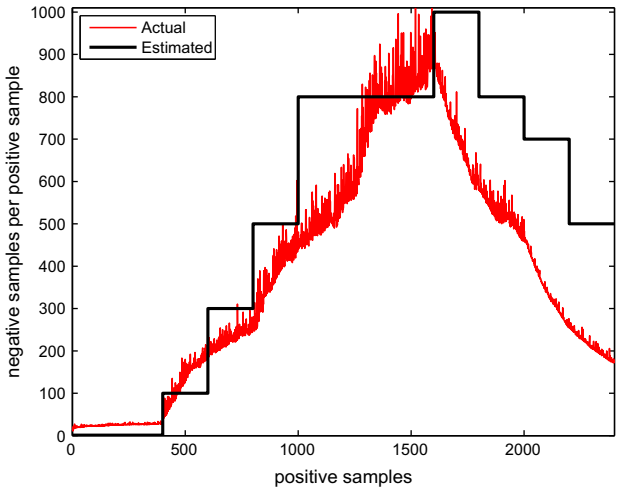


Fig. 9. Class imbalance level adaptation over time. Bold black line is the estimated class imbalance level, while the red curve is the actual class imbalance. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

the second data block ($b = 2$) with the adaptive approach are closer to the target 5%. Performance metrics are also superior with the adaptive approach, as indicated by the F_1, precision and accuracy values. It is also worth noting that *tpr* values are comparable on both approaches, however, the SSBC provided better *fpr* rates, which explain the better performance with other metrics.

5. Experimental validation – face recognition in video surveillance

To validate the proposed SSBC approach, this section presents the methodology and results for experiments with a real-world FR problem. Person re-identification from faces captured in surveillance videos involves detecting individuals of interest that have previously been captured and archived with other video sequences [3,4].

5.1. System for FR in video surveillance

In this scenario, a single IP surveillance camera continuously feeds video frames to a FR system as in Fig. 10. First, the face detection isolates the facial regions of interest (ROIs) in successive frames. Then, the feature extraction module extracts specific characteristics for parallel tracking and recognition streams. Feature extraction, produces the feature vectors **a** for tracking, and **b** for recognition. The face tracking is typically initialized when a different ROI is detected far from others. Tracking follows the position or expression of faces across video frames, whereas the classification function compares the feature patterns to the face models of individuals enrolled to the system. Once a ROI is positively matched against a facial model in the database, the system has potentially detected an individual of interest in video feeds. For robust spatio-temporal FR, the face tracker allows to regroup facial captures for different people, and accumulate the recognition scores according to facial trajectories.

Feature extraction for the recognition stream transforms the ROIs to obtain 32 discriminant features. Initially, pixel intensity

Table 7
Average performance measures for the BC with random under sampling (RUS) and the proposed SSBC for the six data blocks ($b = 1$ to $b = 6$). The standard deviation is detailed between parenthesis.

Approach	Measure	$b = 1$	$b = 2$	$b = 3$	$b = 4$	$b = 5$	$b = 6$
BC w/RUS	*fpr*	5.84%	5.85%	5.85%	5.84%	5.84%	5.86%
		(0.878)	(0.745)	(0.701)	(0.719)	(0.717)	(0.723)
	tpr	58.08%	57.50%	59.08%	58.83%	58.33%	58.05%
	Recall	(3.575)	(4.861)	(3.262)	(3.604)	(3.518)	(3.436)
	Precision	25.11%	3.19%	1.67%	1.05%	2.45%	7.67%
		(2.271)	(0.272)	(0.141)	(0.085)	(0.214)	(0.587)
	F_1	0.350	0.060	0.032	0.021	0.047	0.135
		(0.023)	(0.005)	(0.003)	(0.002)	(0.004)	(0.009)
	acc	93.00%	94.03%	94.09%	94.12%	94.07%	93.84%
		(0.802)	(0.731)	(0.696)	(0.716)	(0.711)	(0.698)
SSBC	*fpr*	5.84%	4.92%	5.01%	5.01%	4.98%	5.01%
		(0.878)	(0.107)	(0.073)	(0.042)	(0.071)	(0.073)
	tpr	58.08%	56.48%	57.77%	57.67%	57.42%	58.12%
	Recall	(3.575)	(2.256)	(4.481)	(3.520)	(2.472)	(2.787)
	Precision	25.11%	3.68%	1.89%	1.18%	2.80%	8.81%
		(2.271)	(0.159)	(0.145)	(0.070)	(0.131)	(0.389)
	F_1	0.350	0.069	0.037	0.023	0.053	0.153
		(0.023)	(0.003)	(0.003)	(0.001)	(0.002)	(0.007)
	acc	93.00%	94.95%	94.93%	94.95%	94.92%	94.68%
		(0.802)	(0.107)	(0.073)	(0.041)	(0.072)	(0.073)

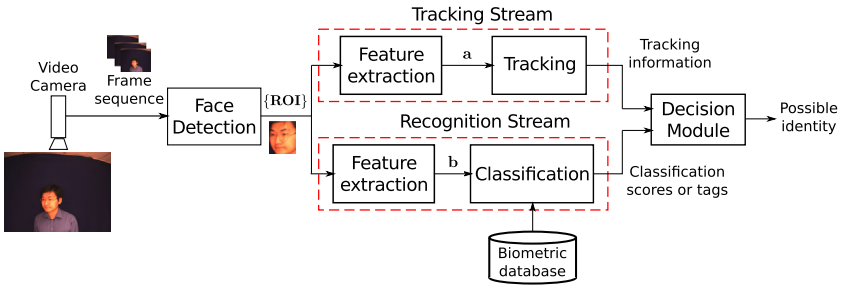


Fig. 10. A generic system for face recognition in video surveillance.

and multi-block local binary pattern (MBLBP) features are concatenated, and the 32 most relevant `f` features are selected through principal component analysis (PCA). Features of unseen faces are then projected against the PCA hyperplanes before classification, and are compared against facial models of target individuals in the classification module to provide a classification score or label. The CAMSHIFT algorithm [48] is used for face tracking according to HSV color histogram features.

Several specialized architectures have been proposed for FR in video surveillance [5]. For instance, the open-set Transduction Confidence Machine-*k*NN (TCM-*k*NN) algorithm [6] modifies the traditional *k*NN, using transduction to calculate a measure of strangeness between target and non-target samples and to provide a reject option for unknown individuals. Ekenel *et al.* [7] combine a *k*-Nearest Neighbor and Gaussian Mixture Modeling with three different metrics to estimate the contribution of individual frames to the overall decision. Kamgar-Parsi *et al.* [8] proposed a morphing approach to generate new synthetic reference data and improve the separability of target and non-target classes.

Since individuals of interest must be detected, systems for FR in video surveillance have effectively been modeled in terms of independent user-specific detection problems, each one implemented using one or more binary (1- or 2-class) classifiers [5]. This modular approach was employed in [9], with user-specific support vector machine classifiers, and in [5] with user-specific ensembles of 2-class ARTMAP neural network classifiers and with responses combined through BC. In this context, the use of a binary ensemble of classifiers per person to improve performance is justified by the limited amount of positive class data for classifier design, and by the uncertainty of facial models with respect to the complexity of real-world video scenes [5].

In this paper, the proposed SSBC approach is evaluated for the adaptation of user-specific ensembles to an imbalanced input data stream in face re-identification applications. Each individual is modeled as an ensemble of 2-class classifiers [5]. Ensembles of classifiers are co-jointly trained using a PSO-based training strategy. It allows for the generation of a diversified pool of ARTMAP neural networks, and trained detectors are then selected and combined in the ROC space using BC.. The approach is compared to a static BC obtained with IBC and where imbalanced data is processed using random under-sampling (RUS) of the negative class and using the one-sided selection (OSS) rule [17] to reduce data dimensionality.

5.2. FIA video data

Video data for this experiment is extracted from the Carnegie Mellon University – Face in Action (FIA) database [15]. This database contains 20 s video sequences for 244 unique individuals simulating a passport checking scenario, captured at a 640×480 pixels resolution at a 30 Hz frame rate. Data is captured over three different capture sessions, each separated by a three months interval.

The capture setup used three camera pairs, with two focal lengths, 2.8 mm (normal) and 4.8 mm (zoomed). Each pair is positioned at different horizontal positions (frontal, left and right).

Fig. 11. Example of the 6 FIA views using the setup capture.

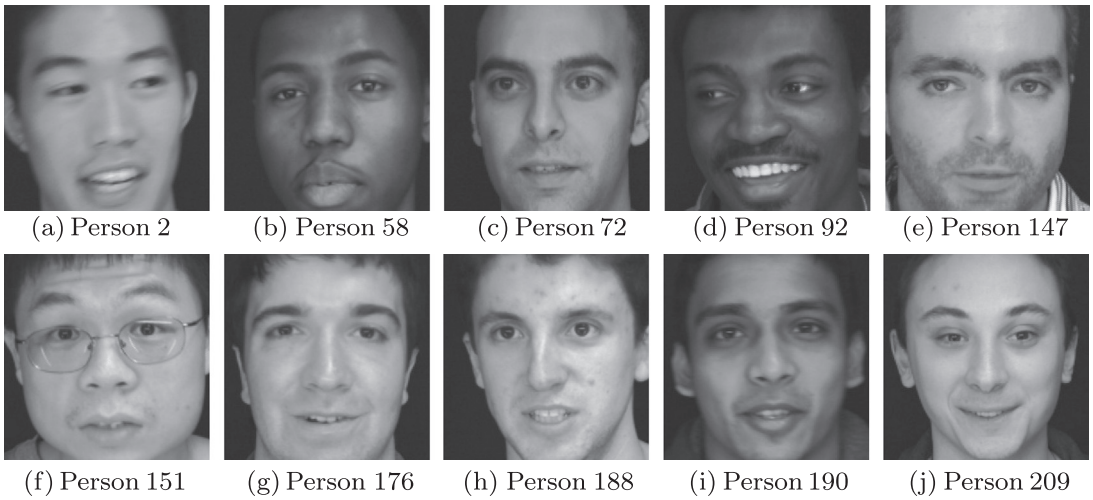


Fig. 12. Target individuals selected in the FIA database. Faces were extracted from the FIA video streams using the Haar cascade based face detector implemented in OpenCV [49].

Thus, each individual is captured from six different views, as shown in Fig. 11. This experiment uses the frontal camera with both focal lengths on all three sessions as the video stream from the single IP camera in Fig. 10.

The initial enrollment process considers a watch list with 10 individuals of interest selected in the database (labeled as person 2, 58, 72, 92, 147, 151, 176, 188, 190 and 209) and depicted in Fig. 12. Each individual is the positive or target class for one detector module (EoC) as described in [5]. The expected class imbalance level is $\lambda = 1/243$, although the actual value depends on the number of facial regions extracted per video sequence.

To improve classifier discrimination of each individual detector module, we use an Universal Background Model (UBM) [50] is built using negative reference samples. Besides the UBM, a cohort model (CM) of the other target individuals is also used to provide negative samples and improve separation between target individuals in the watch list. Individuals in the data base are split in two for training and test. For each individual in the watch list, 100 negative class individuals are selected for training (from UBM and CM [50]), and other 100 negative class individuals are selected for testing, with maximum class imbalance $\lambda^{max} = 1/100$.

5.3. Experimental protocol

Experiments with the FIA video data follows the diagram shown in Fig. 13. It represents the operation for one target individual (thus, this is replicated 10 times, one for each target individual). Initially, a pool of diversified classifiers C is generated from training data using a DPSO training strategy to co-jointly optimize all parameters of a PFAM neural network [16] with seven sub-swarms. Training and validation subsets are selected to generate the initial pool of PFAM neural network classifiers. At the end of the optimization process, the local best classifier from each DPSO sub-swarm is selected for the initial pool of diversified classifiers

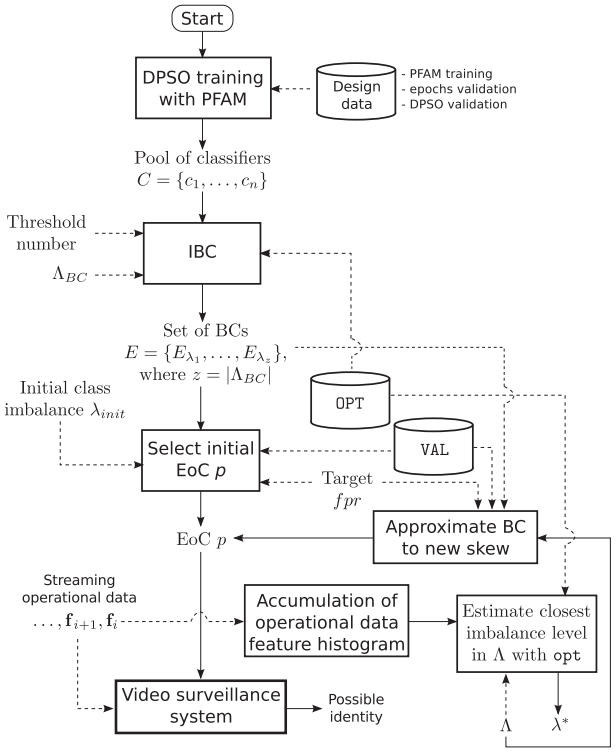


Fig. 13. Flowchart diagram of experimental protocol.

C to perform IBC. This pool of classifiers is then used to optimize a set of BCs E using IBC, one for each class imbalance level in Λ_{BC}. An initial operational point p is selected for a class imbalance level $\lambda_{init} = 1/1$ and a target $fpr = 1\%$.

During operation, the video surveillance system (as in Fig. 10) uses the selected EoC p to evaluates the stream of operational data to identify the target individual in the current frame. This module also tracks face using a CAMSHIFT [48] tracker to follow the movement and location of different faces over time, and to verify if the same individual is appearing in consecutive video frames. Tracked trajectories are used for reliable recognition, by accumulating the positive predictions of each detector module over several frames. If during the last 30 frames (one second) the same individual in the video stream has been detected at least $t_{det} = 8$ times, the system outputs a positive recognition for the target individual. Otherwise, the person is considered to be a non-target.

At the same time, the system accumulates operational data over time for the last 30 min to estimate the level of class imbalance λ^*. After the current elapsed time is higher than the update time $t_u = 15$, the operational data level of closest class imbalance λ^* in Λ is estimated with `opt` as the reference data. Then, the BC is approximated to λ^* and an operational point is selected for the target $fpr = 1\%$ to update p for operation.

Operational test data is extracted from sessions 2 and 3 of the FIA database, using both focal lengths. These four parts of the original database have their video sequences split in two equal parts of 10 s, to produce 8 blocks of video data with one target individual and 100 non-target individuals (CM and unseen individuals). Each of these blocks are used to simulate 30 min of time.

During the test, the level of class imbalance in the test data changes over time in these 8 blocks of 30 min, with the following levels of class imbalance sequence: 1/20,1/35,1/100,1/65,1/100, 1/80,1/60 and 1/15. These values are related to the number of unique individuals per 30 min block. Thus, the actual class imbalance level is known only after extracting the facial regions. The simulation evaluates the system for a total of 4 h of operations.

Table 8
Doddington's zoo analysis adapted for a sequence of binary decisions over a video track to decide the individual identity. False rejection rate (frr) and false positive rate (fpr) thresholds are applied to each individual detector module.

Category	Positive class	Negative class
Sheep	$frr < 50\%$ and not a lamb	$fpr \leqslant 30\%$
Lamb	At least 3% of non-target individuals are wolves	–
Goat	$frr \geqslant 50\%$ and not a lamb	–
Wolf	–	$fpr > 30\%$

The experiment assumes a maximum class imbalance level to adapt $\lambda^{max} = 1/100$. Class imbalance changes in `test` are achieved by randomly removing individuals from each 30 min block. For operation, the experiment uses $\Lambda = \{1/1, 1/10, 1/20, 1/30, \ldots, 1/100\}, \Lambda_{BC} = \{1/1, 1/10, 1/50, 1/100\}$, target $fpr = 1\%$ and $\lambda^{init} = 1/1$. The Class imbalance level closest to the known levels in Λ_{BC} is estimated every $t_u = 15$ minutes, over the last 30 min interval. The experiments are replicated 10 times using 2×6-fold cross-validation to generate training data. After replication five, the 6 folds are randomly regenerated for the next five replications.

FIA Face images from video sequences in session one, captured with both the 2.8 mm and 4.8 mm frontal cameras, are used to generate the pool of diversifier classifiers C and define the BCs for the initial class imbalances in Λ_{BC}. A total of 120 facial samples per individual are randomly selected from both focal distances to build a system design data set `D` with $\lambda = 1/100$. The `D` data set is divided in 6 folds, each fold with $\lambda^{max} = 1/100$. For system operation, frontal video sequences from FIA in sessions 2 and 3 are used with both focal distances to build the `test` data set. Each 20 s original video sequence is split in two 10 s video sequence, providing 8 different video sequences for each individual. Each test video sequence is used only once in one time interval of this experiment.

Training data folds in `D` are split in six folds as follows. The $\mathtt{D}_t^t$ uses 2 folds, with a total of 40 positive samples. Each of the remainder data sets uses one fold, each with 20 positive samples. The $\mathtt{D}_t^e$ validation data is used to stop the number of training epochs and avoid over-fitting, whereas $\mathtt{D}_t^f$ is the validation data to evaluate the fitness function of the DPSO learning strategy. The negative data in sets $\mathtt{D}_t^t, \mathtt{D}_t^e$ and $\mathtt{D}_t^f$ are balanced through random under-sampling for classifier training. The data set $\mathtt{D}_t^o$ is used to generate the set of data sets `OPT` following Λ_{BC}, while $\mathtt{D}_t^v$ is used to create `OPT`. Each fold has 2000 negative samples from the cohort and universal background models, providing class imbalance levels up to $\lambda = 1/100$.

To compare the proposed SSBC, two under sampling BC approaches are evaluated. The first uses random under sampling (RUS) to balance the data used to optimize the BC (`opt`) and select the operational point (`val`), while the second uses the one-sided selection (OSS) rule. The pool of classifier C is generated as with the proposed adaptive approach, however, only one BC is optimized to select a single operational point for the entire simulation.

During operation, test data is streamed over time and is evaluated by the system. This data is accumulated by the adaptive BC approach, to estimate the closest class imbalance level in Λ_{BC} at every 15 min, over the last 30 min data. To define the closest class proportions, $b = \lfloor \sqrt{L_+} \rfloor = 4$ bins per feature are used, where $L_+ = 20$ is the positive class cardinality of the labeled reference data.

The system is evaluated regarding three different analysis types. Transactional analysis considers only matching performance, based on traditional metrics: tpr ($recall$), fpr, $precision$ and F_1. Time based analysis evaluates overall performance over time,

Table 9
Average performance measures for different approaches for a target $fpr = 1\%$ on test segments at different update times $t = 1, 2, \ldots, 8$ times. BC with RUS and OSS are static approaches, using, respectively, random under sampling (RUS) and the one side selection (OSS) sample selection rules to improve training with imbalanced data. The standard deviation value is shown in parenthesis. Bold values indicate the best results.

Approach	Measure	Update period							
		$t = 1$	$t = 2$	$t = 3$	$t = 4$	$t = 5$	$t = 6$	$t = 7$	$t = 8$
SSBC	*fpr*	**4.89%**	**1.20%**	**1.65%**	**1.85%**	**1.16%**	**1.09%**	**0.66%**	**0.70%**
		(0.024)	(0.008)	(0.008)	(0.012)	(0.006)	(0.008)	(0.005)	(0.006)
	tpr	65.58%	49.66%	54.53%	55.67%	53.42%	51.00%	47.52%	49.85%
	Recall	(0.299)	(0.329)	(0.247)	(0.308)	(0.261)	(0.306)	(0.394)	(0.399)
	Precision	43.68%	55.09%	41.15%	45.33%	41.99%	47.17%	53.59%	67.93%
		(0.225)	(0.315)	(0.209)	(0.187)	(0.177)	(0.198)	(0.335)	(0.314)
	F_1	**0.492**	**0.518**	**0.446**	**0.479**	**0.450**	**0.470**	**0.498**	**0.550**
		(0.217)	(0.255)	(0.187)	(0.212)	(0.191)	(0.221)	(0.332)	(0.344)
BC w/ RUS	*fpr*	4.89%	4.32%	5.82%	5.93%	4.65%	4.57%	3.45%	3.63%
		(0.024)	(0.021)	(0.025)	(0.027)	(0.025)	(0.025)	(0.020)	(0.024)
	tpr	65.58%	67.40%	69.71%	69.87%	69.01%	66.06%	61.68%	64.02%
	Recall	(0.299)	(0.292)	(0.186)	(0.231)	(0.153)	(0.241)	(0.320)	(0.319)
	Precision	43.68%	38.94%	23.37%	29.23%	23.93%	27.04%	34.25%	54.43%
		(0.225)	(0.211)	(0.127)	(0.109)	(0.107)	(0.108)	(0.184)	(0.237)
	F_1	0.492	0.470	0.319	0.382	0.332	0.349	0.414	0.550
		(0.217)	(0.195)	(0.136)	(0.113)	(0.129)	(0.134)	(0.212)	(0.237)
BC w/ OSS	*fpr*	2.14%	2.04%	2.19%	2.49%	1.63%	1.70%	1.16%	1.49%
		(0.013)	(0.011)	(0.007)	(0.013)	(0.007)	(0.008)	(0.005)	(0.008)
	tpr	51.51%	53.94%	53.38%	52.32%	51.53%	49.95%	51.19%	53.47%
	Recall	(0.325)	(0.294)	(0.227)	(0.280)	(0.249)	(0.315)	(0.364)	(0.362)
	Precision	54.32%	49.96%	35.09%	38.46%	35.52%	37.77%	49.90%	63.44%
		(0.297)	(0.278)	(0.177)	(0.154)	(0.181)	(0.192)	(0.281)	(0.273)
	F_1	0.497	0.487	0.387	0.407	0.392	0.393	0.474	0.540
		(0.276)	(0.229)	(0.181)	(0.173)	(0.199)	(0.220)	(0.303)	(0.301)

and uses the accumulation strategy that relies on a face tracker. Individuals are identified only after accumulating a minimum number of positive detections over a face track, i.e., after surpassing a decision threshold $t_{det} = 8$ detections in the video stream over 30 consecutive frames (1 s).

Finally, subject based analysis investigates the system performance for each unique individual, using the Doddington's Zoo taxonomy [18,19]. Doddington's Zoo analysis categorizes individuals in one of the following four categories: (1) sheeps, easy to identify individuals (positive or negative class), (2) goats, positive class individuals that are difficult to identify, (3) wolves, negative class individuals that impersonate one or more positive class individuals or (4) lambs, positive class individuals that are easy to impersonate.

The traditional way to define the system users likeliness in a Doddington zoo's category is through the classifier output scores for all tested samples [19]. However, BC techniques provide crisp decisions and the confusion matrix of individual accumulated decisions is used to categorize individuals according to Table 8. This approach is based on the technique used in [6], and considers the *fpr* and the false rejection rate (*frr*, the percentage of positive individual rejections) for one individual detector module. The *frr* and *fpr* decision thresholds were selected to provide a more conservative categorization.

5.4. Results and discussion

Table 9 details the mean transactional performance for the compared approaches, as well as standard deviation values (between parenthesis). At time $t = 1$, both the random under sampling static approach and the proposed adaptive approach uses the same EoC, optimized for balanced data. The BC with OSS based approach uses different under-sampled data. After time $t = 2$ the SSBC approach uses data in the last 30 min to estimate the class imbalance and approximate a new BC of classifiers. The SSBC technique selects an operational point closer to the target $fpr = 1\%$,

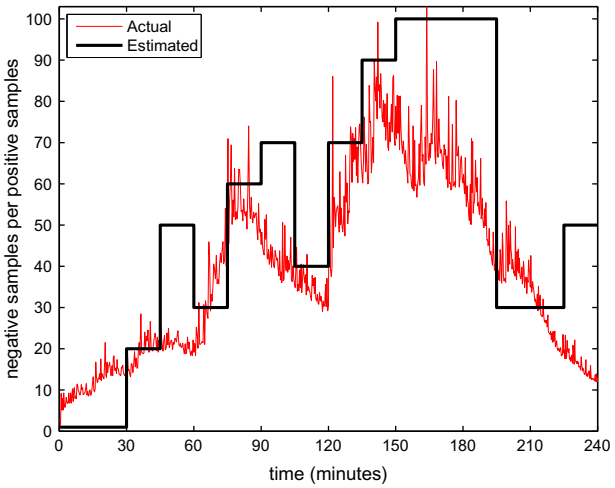


Fig. 14. Hellinger distance based estimation of class imbalance level to the closest level in Λ_{BC}.

with smaller standard deviation values. On the other hand, the BC with RUS approach selects an operational point with higher *fpr* values, while the BC with OSS approach provides lower *fpr* values, at the expense of low positive class performance. This is a problem for detecting target individuals, as demonstrated later in Table 12. Surveillance applications are challenged by severe skewed data. The proposed SSBC technique reduces the number of false positive detections and keeps high positive performance, thereby providing better support for a human operator.

The SSBC technique relies on the Hellinger distance to estimate the closest adjacent class imbalance in Λ_{BC}. Fig. 14 details the estimations for one of the replications. The bold line is the estimated class imbalance level from the labeled `OPT` data sets and unlabeled test data. The other curve indicate the actual class imbalance level perceived in the test data.

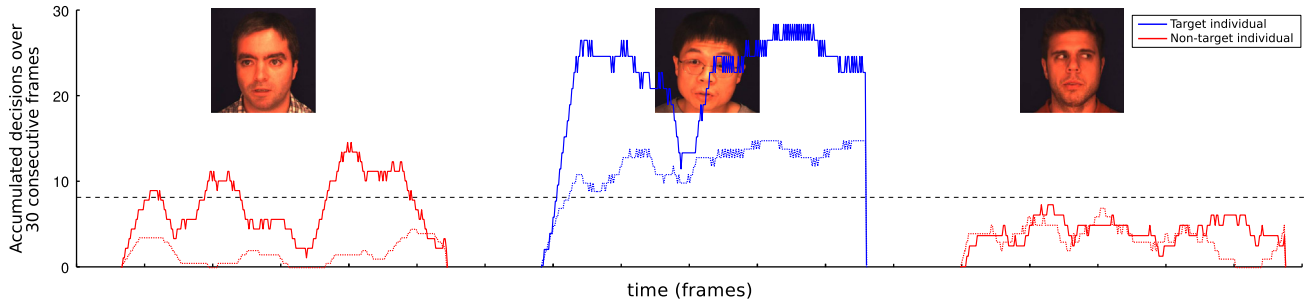


Fig. 15. Time analysis for the module of individual 151. The blue line is for a target individual of interest, while the red line is the typical accumulation of a non-target individual. Solid lines are the BC with RUS approach, while dotted lines are the proposed SSBC. Positive detections are made after accumulating at least 8 positive detections over a sliding window of 30 consecutive video frames. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Time analysis is performed based over positive individual detections, using a CAMSHIFT face tracker to accumulate decisions over time to provide a more reliable decision. Once a face is found in the video sequence, its location is used to initialize a tracker that follows the face until it leaves the scene. Every time a face is found in the same area followed by the tracker, the individual detection is accumulated over a sliding window of 30 frames. When the accumulated positive detections reach the threshold $t_{det} = 8$, a decision is made and the target individual is identified.

Fig. 15 compares the accumulation process of SSBC and BC with RUS approaches. The blue line is the accumulation for a target individual in the watch list, while the red lines are the typical accumulation of a non-target individual (outside the watch list). The black horizontal dashed line is the detection threshold t_{det}. Accumulating decisions over time allows better decisions, as incorrect positive detections do no represent identity decisions. Performance wise, the proposed SSBC approach (dashed lines) perform better, generating ensembles that are more resistant to false positive detections. In the example, one individual outside the watch list is identified incorrectly when using the BC with RUS approach. This situation is not repeated with SSBC, that provides a *fpr* value close to the target 1% and accumulates positive decisions slower than the EoC with RUS approach.

Tables 10–12 details the time analysis based individual detection confusion matrix for the SSBC, BC with RUS and BC with OSS approaches. The main diagonal (bold numbers) represents the identification performance of individuals targeted in the watch list. Performance is measured as the proportions of detections over all appearances in the video sequences (not individual frames). For target individuals (main diagonal) a higher value is expected, while for non-target individuals the system must provide a lower value.

The SSBC provides the best performance trade off between high performance on target individuals and resistance to false negatives on non-targets, which is related to the higher F_1 scores obtained in Table 9. The random under sampling BC improves the positive class performance, but at the expense of a decrease on the non-target

Table 10
Positive identification confusion matrix for the SSBC approach. Each column is the detector module for each target individual in the watch list, while lines are the individuals detected in the video sequences. Each line represents the target individuals in the watch list and columns are the detectors for each target individual. Color codes represent the individuals according to the Dodington's zoo analysis. For each module, green cells are sheep like positive individuals, blue are goat like positive individuals, yellow are lamb like individuals, red are wolf like negative individuals and white are sheep like negative individuals.

Indiv.	MOD_2 (%)	MOD_{58} (%)	MOD_{72} (%)	MOD_{92} (%)	MOD_{147} (%)	MOD_{151} (%)	MOD_{176} (%)	MOD_{188} (%)	MOD_{190} (%)	MOD_{209} (%)
2	**60.42**	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
58	1.04	**59.90**	0.52	0.00	0.00	0.00	0.00	0.00	1.04	0.00
72	0.00	0.00	**59.38**	0.00	3.12	0.00	0.00	1.04	0.00	34.90
92	4.69	2.60	0.00	**68.75**	10.42	4.69	0.00	1.56	4.17	0.00
147	0.00	0.00	2.60	0.00	**33.33**	0.52	0.00	1.04	0.52	4.69
151	0.00	0.00	0.00	0.00	0.00	**70.31**	9.38	2.60	0.00	0.00
176	0.00	0.00	0.00	0.00	0.00	0.00	**46.35**	9.38	0.52	0.00
188	0.00	0.00	0.52	0.00	3.12	8.33	10.42	**98.44**	0.00	4.17
190	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	**79.69**	0.00
209	0.00	0.00	20.31	0.00	0.52	0.00	0.00	0.00	0.00	**98.96**

Table 11
Positive identification confusion matrix for the random under sampling BC. The structure is the same as in Table 10.

Indiv.	MOD_2 (%)	MOD_{58} (%)	MOD_{72} (%)	MOD_{92} (%)	MOD_{147} (%)	MOD_{151} (%)	MOD_{176} (%)	MOD_{188} (%)	MOD_{190} (%)	MOD_{209} (%)
2	**81.77**	1.04	3.12	1.56	0.52	3.65	0.00	1.04	1.56	1.56
58	3.12	**75.52**	11.98	1.56	0.00	0.52	0.00	0.00	23.96	0.00
72	0.00	0.00	**74.48**	0.52	10.94	0.00	0.00	5.21	6.77	68.23
92	8.33	13.02	0.52	**72.92**	17.19	12.50	0.00	1.56	14.58	0.00
147	0.00	0.00	11.98	0.00	**48.96**	1.04	0.00	5.73	4.69	9.90
151	0.00	0.00	0.00	0.00	0.00	**79.17**	22.40	10.42	1.04	4.17
176	5.73	0.00	0.00	0.00	3.65	2.60	**60.94**	23.44	2.60	0.00
188	0.00	0.00	2.08	0.00	9.38	19.79	28.65	**94.79**	0.00	9.38
190	0.00	8.33	1.56	0.52	0.52	0.52	0.00	0.00	**96.88**	0.52
209	0.00	0.00	35.94	0.00	2.60	2.08	0.00	0.00	0.00	**99.48**

Table 12
Positive identification confusion matrix for the BC with OSS. The structure is the same as in Table 10.

Indiv.	MOD_2 (%)	MOD_{58} (%)	MOD_{72} (%)	MOD_{92} (%)	MOD_{147} (%)	MOD_{151} (%)	MOD_{176} (%)	MOD_{188} (%)	MOD_{190} (%)	MOD_{209} (%)
2	**53.65**	0.00	0.00	0.00	0.00	0.00	0.52	0.00	0.00	0.52
58	0.00	**52.08**	1.04	1.04	0.00	0.00	0.52	0.00	11.46	0.00
72	0.00	0.00	**52.60**	0.00	3.65	0.00	0.00	0.00	1.04	34.90
92	1.04	9.38	0.00	**59.90**	9.38	3.65	0.00	0.00	0.00	0.00
147	0.00	0.00	4.17	0.00	**33.33**	0.00	0.52	0.00	3.12	6.77
151	0.00	0.00	0.00	0.00	0.00	**69.79**	7.81	3.65	0.00	1.56
176	1.04	0.00	0.00	0.00	0.00	0.52	**31.25**	3.65	0.00	0.00
188	0.00	0.00	0.00	0.00	4.17	13.02	30.73	**97.92**	0.00	3.65
190	0.00	0.00	1.56	0.00	0.00	0.00	0.00	0.00	**75.52**	0.00
209	0.00	0.00	28.65	0.00	0.00	0.00	0.00	0.00	0.00	**92.71**

individuals. Finally, the resistance to false negatives of BC with OSS is similar to that of SSBC, but performance on the positive class is inferior to that provided by SSBC.

A Dodington's zoo subject-based analysis in Tables 10–12 indicated through colors in these tables. Different thresholds are used to make a decision regarding Dodington's zoo categories as indicated in Table 8. For these thresholds, the proposed SSBC approach provides the best performance for the positive and negative classes. For the target individuals, there is one lamb like individual and 2 goat like individuals, for a total of 12 wolf like individuals in the non-target class (100 individuals).

Whereas the confusion matrix shows individual frames detection rate improvements for the positive class when using the random under sampling BC, a total of six lamb like individuals and one goat like individual are observed with this approach. These six lamb like individuals are related to the higher *fpr* rates in Table 9, and a total 38 wolf like individuals are identified (three times more than SSBC). Finally, the BC with OSS static approach has no lamb like individuals, two goats and 12 wolves. Whereas having no goat like individuals, overal positive class performance of BC with OSS is consistently bellow that obtained by the BC with SSBC for all individuals. This result is related to the higher *tpr* and closer value to *fpr* associated to BC with SSBC in Table 9.

6. Conclusions

EoCs have been proposed in the literature to reduce the impact on performance from imbalanced class distributions. The BC of ensembles in 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 with imbalanced data that to corresponds to that of operational conditions produces a better combination of base classifiers. In this paper, an adaptive system based in SSBC is proposed to select the most accurate BC of base classifiers according to the target fpr and class imbalance. Imbalanced data is used to generate several BCs in the decision space, 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 estimate the most accurate BC of classifiers. Instead of full re-calculation of BCs, 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, and maintaining a comparable level of performance.

Experiments using real-world FIA video streams for face re-identification were performed to compare the adaptive SSBC approach to static BC obtained with the RUS approach and the OSS rule. Results indicate the advantages of adapting the BC to the operation class imbalance, which provides a good trade-off between positive (target individuals) and negative (non-target individuals) classes. Transaction based analysis indicate the operation *fpr* values remain similar to the desired values when using the proposed adaptive approach, while providing consistently higher F_1 measures. In time-based analysis, where positive classification prediction are accumulated over a window of time, the proposed adaptive BC allow to accurately re-identify individuals of interest.

SSBC depends heavily on the granularity of the pre-trained λ levels, affecting a trade-off between accuracy and resources to store BC curves and reference data. Future research will focus on improving approximations of BC to estimated imbalance levels. The technique currently selects ensembles from the original BCs optimized for the adjacent class imbalance levels with data following the new class imbalance level. However, this approximation may benefit from a strategy to combine vertices in the ROCCH by normalizing the proportions associated to each adjacent class imbalance level in regard to the new class imbalance level. Another modification is to use training data with several levels of class imbalance to generate the pool of diversified classifiers to perform BC and avoid the limitation associated to the number of original BC curves and the quality of approximation.

Acknowledgement

This work was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC), and Public Security Technical Program (PTSP 03-0401BIOM) of the Defense Research and Development Canada, Centre for Security Science (CSS).

References

[1] M. Galar, A. Fernandez, E. Barrenechea, H. Bustince, F. Herrera, A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches, IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews 42 (4) (2012) 463–484, http://dx.doi.org/10.1109/TSMCC.2011.2161285.

[2] M. Fischer, H. Ekenel, R. Stiefelhagen, Person re-identification in tv series using robust face recognition and user feedback, Multimedia Tools and Applications 55 (1) (2011) 83–104, http://dx.doi.org/10.1007/s11042-010-0603-2.

[3] O. Hamdoun, F. Moutarde, B. Stanciulescu, B. Steux, Person re-identification in multi-camera system by signature based on interest point descriptors collected on short video sequences, in: Second ACM/IEEE International Conference on Distributed Smart Cameras, 2008, pp. 1–6, doi:http://dx.doi.org/10.1109/ICDSC.2008.4635689.

[4] R. Satta, G. Fumera, F. Roli, Fast person re-identification based on dissimilarity representations, Pattern Recognition Letters 33 (14) (2012) 1838–1848, http://dx.doi.org/10.1016/j.patrec.2012.03.026.

[5] C. Pagano, E. Granger, R. Sabourin, D.O. Gorodnichy, Detector ensembles for face recognition in video surveillance, in: Proceedings of the 2012 International Joint Conference on Neural Networks, 2012, pp. 1–8, doi:http://dx.doi.org/10.1109/IJCNN.2012.6252659.

[6] F. Li, H. Wechsler, Open set face recognition using transduction, IEEE Transactions on Pattern Analysis and Machine Intelligence 27 (11) (2005) 1686–1697, http://dx.doi.org/10.1109/TPAMI.2005.224.

[7] H.K. Ekenel, J. Stallkamp, R. Stiefelhagen, A video-based door monitoring system using local appearance-based face models, Computer Vision and Image Understanding 114 (5) (2010) 596–608, http://dx.doi.org/10.1016/j.cviu.2009.06.009.

[8] B. Kamgar-Parsi, W. Lawson, B. Kamgar-Parsi, Toward development of a face recognition system for watchlist surveillance, IEEE Transactions on Pattern Analysis and Machine Intelligence 33 (2011) 1925–1937, http://dx.doi.org/10.1109/TPAMI.2011.68.

[9] H.K. Ekenel, L. Szasz-Toth, R. Stiefelhagen, Open-set face recognition-based visitor interface system, in: Proceedings of the 7th International Conference on Computer Vision Systems, Springer-Verlag, Berlin, Heidelberg, 2009, pp. 43–52. doi:http://dx.doi.org/10.1007/978-3-642-04667-4_5.

[10] W. Khreich, E. Granger, A. Miri, R. Sabourin, Iterative boolean combination of classifiers in the ROC space: an application to anomaly detection with HMMs, Pattern Recognition 43 (8) (2010) 2732–2752, http://dx.doi.org/10.1016/j.patcog.2010.03.006.

[11] T. Fawcett, An introduction to ROC analysis, Pattern Recognition Letters 27 (8) (2006) 861–874, http://dx.doi.org/10.1016/j.patrec.2005.10.010.

[12] T. Landgrebe, P. Paclik, R. Duin, A. Bradley, Precision-recall operating characteristic (p-roc) curves in imprecise environments, in: Proceedings of the 18th International Conference on Pattern Recognition, 2006, pp. 123–127, doi:http://dx.doi.org/10.1109/ICPR.2006.941.

[13] J. Davis, M. Goadrich, The relationship between precision-recall and ROC curves, in: Proceedings of the 23rd International Conference on Machine Learning, New York, NY, USA, 2006, pp. 233–240, doi:http://dx.doi.org/10.1145/1143844.1143874.

[14] V. González-Castro, R. Alaiz-Rodríguez, L. Fernández-Robles, R. Guzmán-Martínez, 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, 2010, pp. 284–293.

[15] R. Goh, L. Liu, X. Liu, T. Chen, The CMU Face in Action (FIA) Database, in: Proceedings of the Second International Workshop on Analysis and Modelling of Faces and Gestures, Berlin, Germany, 2005, pp. 255–263, doi:http://dx.doi.org/10.1007/11564386_20.

[16] J.-F. Connolly, E. Granger, R. Sabourin, Evolution of heterogeneous ensembles through dynamic particle swarm optimization for video-based face recognition, Pattern Recognition 45 (7) (2012) 2460–2477, http://dx.doi.org/10.1016/j.patcog.2011.12.016.

[17] M. Kubat, S. Matwin, Addressing the curse of imbalanced training sets: one-sided selection, in: Proceedings of the Fourteenth International Conference on Machine Learning, 1997, pp. 179–186.

[18] G. Doddington, W. Liggett, A. Martin, M. Przybocki, D. Reynolds, Sheep, goats, lambs and wolves: a statistical analysis of speaker performance in the NIST 1998 speaker recognition evaluation, in: International Conference on Spoken Language Processing, 1998.

[19] A. Rattani, G. Marcialis, F. Roli, An experimental analysis of the relationship between biometric template update and the Doddington's Zoo: a case study in face verification, in: P. Foggia, C. Sansone, M. Vento (Eds.), Image Analysis and Processing – ICIAP 2009, Lecture Notes in Computer Science, vol. 5716, Springer, Berlin/Heidelberg, 2009, pp. 434–442. doi:http://dx.doi.org/10.1007/978-3-642-04146-4_47.

[20] P.V.W. Radtke, E. Granger, R. Sabourin, D. Gorodnichy, Adaptive selection of ensembles for imbalanced class distributions, in: Proceedings of the 21st International Conference of Pattern Recognition, 2012, pp. 1–5.

[21] X. Guo, Y. Yin, C. Dong, G. Yang, G. Zhou, On the class imbalance problem, in: Proceedings of the 4th International Conference on Natural Computation, 2008, pp. 192–201, doi:http://dx.doi.org/10.1109/ICNC.2008.871.

[22] Y. Sun, A.K.C. Wong, M.S. Kamel, Classification of imbalanced data: a review, International Journal of Pattern Recognition and Artificial Intelligence 23 (4) (2009) 687–718, http://dx.doi.org/10.1142/S0218001409007326.

[23] S. Lawrence, I. Burns, A. Back, A. Tsoi, C. Giles, Neural network classification and prior class probabilities, in: G. Orr, K.-R. Müller (Eds.), Neural Networks: Tricks of the Trade, Lecture Notes in Computer Science, vol. 1524, Springer, Berlin/Heidelberg, 1998, p. 545. doi:http://dx.doi.org/10.1007/3-540-49430-8_15.

[24] S.B. Kotsiantis, P.E. Pintelas, Mixture of expert agents for handling imbalanced data sets, Annals of Mathematics, Computing & Teleinformatics 1 (1) (2003) 46–55.

[25] P.E. Hart, The condensed nearest neighbor rule (corresp.), IEEE Transactions on Information Theory 14 (3) (1968) 515–516, http://dx.doi.org/10.1109/TIT.1968.1054155.

[26] N.V. Chawla, K.W. Bowyer, L.O. Hall, W.P. Kegelmeyer, SMOTE: synthetic minority over-sampling technique, Journal of Artificial Intelligent Research 16 (1) (2002) 321–357.

[27] T.G. Dietterich, Ensemble learning, in: M.A. Arbib (Ed.), The Handbook of Brain Theory and Neural Networks, second ed., MIT Press, 2002, pp. 405–408.

[28] S. Tulyakov, S. Jaeger, V. Govindaraju, D. Doermann, Review of classifier combination methods, in: S. Marinai, H. Fujisawa (Eds.), Machine Learning in Document Analysis and Recognition, Studies in Computational Intelligence, vol. 90, Springer, Berlin/Heidelberg, 2008, pp. 361–386. doi:http://dx.doi.org/10.1007/978-3-540-76280-5_14.

[29] L. Breiman, Bagging predictors, Machine Learning 24 (1996) 123–140, http://dx.doi.org/10.1023/A:1018054314350.

[30] Y. Freund, R.E. Schapire, A decision-theoretic generalization of on-line learning and an application to boosting, Journal of Computer and System Sciences 55 (1) (1997) 119–139, http://dx.doi.org/10.1006/jcss.1997.1504.

[31] T.K. Ho, The random subspace method for constructing decision forests, IEEE Transactions on Pattern Analysis and Machine Intelligence 20 (8) (1998) 832–844, http://dx.doi.org/10.1109/34.709601.

[32] W. Fan, S.J. Stolfo, J. Zhang, P.K. Chan, AdaCost: misclassification cost-sensitive boosting, in: Proceedings of the Sixteenth International Conference on Machine Learning, ICML '99, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1999, pp. 97–105.

[33] Y. Sun, M.S. Kamel, A.K. Wong, Y. Wang, Cost-sensitive boosting for classification of imbalanced data, Pattern Recognition 40 (12) (2007) 3358–3378, http://dx.doi.org/10.1016/j.patcog.2007.04.009.

[34] N.V. Chawla, A. Lazarevic, L.O. Hall, K.W. Bowyer, SMOTEBoost: improving prediction of the minority class in boosting, in: Proceedings of the 7th European Conference on Principles and Practice of Knowledge Discovery in Databases, 2003, pp. 107–119.

[35] S. Wang, X. Yao, Diversity analysis on imbalanced data sets by using ensemble models, in: IEEE Symposium on Computational Intelligence and Data Mining, 2009, CIDM '09, 2009, pp. 324–331, doi:http://dx.doi.org/10.1109/CIDM.2009.4938667.

[36] D. Tao, X. Tang, X. Li, X. Wu, Asymmetric bagging and random subspace for support vector machines-based relevance feedback in image retrieval, IEEE Transactions on Pattern Analysis and Machine Intelligence 28 (7) (2006) 1088–1099, http://dx.doi.org/10.1109/TPAMI.2006.134.

[37] M.D. la Torre, E. Granger, P.V.W. Radtke, R. Sabourin, D. Gorodnichy, Incremental update of biometric models in face-based video surveillance, in: Proceedings of the 2012 International Joint Conference on Neural Networks, 2012, pp. 1–8, doi:http://dx.doi.org/10.1109/IJCNN.2012.6252658.

[38] C.J.V. Rijsbergen, Information Retrieval, second ed., Butterworth-Heinemann, 1979.

[39] P.V.W. Radtke, T. Wong, R. Sabourin, Solution over-fit control in evolutionary multiobjective optimization of pattern classification systems, International Journal of Pattern Recognition and Artificial Intelligence 23 (06) (2009) 1107–1127, http://dx.doi.org/10.1142/S0218001409007466.

[40] E.M. Dos Santos, R. Sabourin, P. Maupin, Overfitting cautious selection of classifier ensembles with genetic algorithms, Information Fusion 10 (2) (2009) 150–162, http://dx.doi.org/10.1016/j.inffus.2008.11.003.

[41] C. Yang, J. Zhou, Non-stationary data sequence classification using online class priors estimation, Pattern Recognition Letters 41 (8) (2008) 2656–2664, http://dx.doi.org/10.1016/j.patcog.2008.01.025.

[42] Z. Zhang, J. Zhou, Transfer estimation of evolving class priors in data stream classification, Pattern Recognition Letters 43 (9) (2010) 3151–3161, http://dx.doi.org/10.1016/j.patcog.2010.03.021.

[43] E. Granger, W. Khreich, R. Sabourin, D.O. Gorodnichy, Fusion of biometric systems using boolean combination: an application to iris-based authentication, International Journal on Biometrics 4 (3) (2012) 291–315, http://dx.doi.org/10.1504/IJBM.2012.047645.

[44] Q. Tao, R. Veldhuis, Threshold-optimized decision-level fusion and its application to biometrics, Pattern Recognition 42 (2009) 823–836.

[45] J. Daugman, Biometric Decision Landscapes – Technical Report UCAM-CL-TR-482, Tech. rep., Universtity of Cambridge, UK, 2000.

[46] S. Haker, W.M. Wells, S.K. Warfield, I.-F. Talos, J.G. Bhagwat, D. Goldberg-Zimring, A. Mian, L. Ohno-Machado, K.H. Zou, 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 Intervention, vol. I, MICCAI'05, Springer-Verlag, Berlin, Heidelberg, 2005, pp. 506–514, doi:http://dx.doi.org/10.1007/11566465_63.

[47] M.J.J. Scott, M. Niranjan, R.W. Prager, Realisable classifiers: improving operating performance on variable cost problems, in: Proceedings of the Brittish Machine Vision Conference, 1998, pp. 306–315.

[48] G.R. Bradski, Computer vision face tracking for use in a perceptual user interface, Intel Technology Journal 2 (Q2) (1998) 12–21.

[49] G. Bradski, The OpenCV Library, Dr. Dobb's Journal of Software Tools.

[50] A. Brew, P. Cunningham, Combining cohort and UBM models in open set speaker detection, Multimedia Tools and Applications 48 (2010) 141–159, http://dx.doi.org/10.1007/s11042-009-0381-x.