

SUMMARIZING THE PERFORMANCES OF A BACKGROUND SUBTRACTION ALGORITHM MEASURED ON SEVERAL VIDEOS

S. Piérard, and M. Van Droogenbroeck

Department of Electrical Engineering and Computer Science, University of Liège, Belgium
{S.Pierard, M.VanDroogenbroeck}@uliege.be

ABSTRACT

There exist many background subtraction algorithms to detect motion in videos. To help comparing them, datasets with ground-truth data such as CDNET or LASIESTA have been proposed. These datasets organize videos in categories that represent typical challenges for background subtraction. The evaluation procedure promoted by their authors consists in measuring performance indicators for each video separately and to average them hierarchically, within a category first, then between categories, a procedure which we name “summarization”. While the summarization by averaging performance indicators is a valuable effort to standardize the evaluation procedure, it has no theoretical justification and it breaks the intrinsic relationships between summarized indicators. This leads to interpretation inconsistencies. In this paper, we present a theoretical approach to summarize the performances for multiple videos that preserves the relationships between performance indicators. In addition, we give formulas and an algorithm to calculate summarized performances. Finally, we showcase our observations on CDNET 2014.

Index Terms—performance summarization, background subtraction, multiple evaluations, CDNET, classification performance

1. INTRODUCTION

A plethora of background subtraction algorithms have been proposed in the literature [1, 2, 3, 4, 5, 6, 7]. They aim at predicting, for each pixel of each frame of a video, whether the pixel belongs to the background, free of moving objects, or to the foreground. Background subtraction algorithms have to operate in various conditions (viewpoint, shadows, lighting conditions, camera jitter, etc). These conditions are covered by different videos in evaluation datasets, such as CDNET [8, 9] or LASIESTA [10].

Measuring the performance on a single video is done at the pixel level, each pixel being associated with a ground-truth class y (either negative c^- for the background, or positive c^+ for the foreground) and an estimated class \hat{y} (the temporal and spatial dependences between pixels are ignored). Performance indicators adopted in the field of background sub-

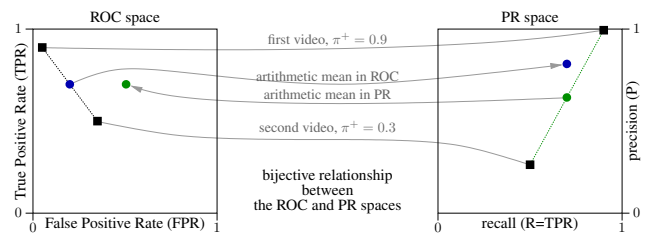


Fig. 1: How can we summarize the performances of a background subtraction algorithm, measured on several videos? This figure shows that arithmetically averaging the performances breaks the bijective relationship that exists between the ROC and Precision-Recall (PR) spaces [11]. This paper presents a new summarization technique that guarantees the coherence between any sets of performance indicators.

traction are mainly those used for two-class crisp classifiers (precision, recall, sensitivity, F-score, error rate, etc).

To compare the behavior of algorithms, it is helpful to “summarize” all the performance indicators that are originally measured on the individual videos. However, to the best of our knowledge, there is no theory for this summarization process, although an attempt to standardize it has been promoted with CDNET. In CDNET, videos are grouped into 11 categories, all videos in a given category having the same importance, and all categories having also the same importance. In other words, the videos are weighted regardless of their size. The standardized summarization process of CDNET is performed hierarchically by computing an arithmetic mean, first between the videos in a category, then between the categories. This averaging is done for seven performance indicators, one by one, independently. While the procedure of CDNET is valuable, it has two major drawbacks related to the interpretability of the summarized values.

First, the procedure breaks the intrinsic relationships between performance indicators. The case of the M4CD algorithm, as evaluated on CDNET, typically illustrates this inconsistency. Despite that the F-score is known to be the harmonic mean of precision and recall, the arithmetic mean of the F-scores across all videos is 0.69 while the harmonic mean between the arithmetically averaged precisions and re-

calls is 0.75. The difficulty to summarize F-scores has already been discussed in [12], when averaging between folds in cross-validation. Such a problem also occurs for other indicators. Moreover, Figure 1 shows that the summarization with the arithmetic mean breaks the bijective relationship between the ROC and PR spaces [11]. It is not equivalent to summarize the performances in one space or the other. In particular, one could obtain a meaningless arithmetically averaged performance point, located in the unachievable part of the PR space [13] (the achievable part is not convex). This leads to difficulties for the interpretation and, eventually, to contradictions between published summarized results.

Second, some indicators (such as the precision, recall, sensitivity, or error rate) have a probabilistic meaning [14]. Arithmetically averaging these indicators does not lead to a value that preserves the probabilistic meaning. This leads to interpretability issues. Strictly speaking, one cannot think in terms of probabilities unless a random experiment is defined very precisely, as shown by Bertrand’s paradox [15].

This paper presents a better summarization procedure that avoids these interpretability issues. In Section 2, we present indicators to measure the performance of an algorithm applied on a single video and pose the random experiment that underpins the probabilistic meaning of these indicators. Then, in Section 3, we generalize this random experiment for several videos, and show that the resulting performance indicators can be computed based only on the indicators obtained separately for each video. In other words, we have established a theoretical model for summarization that guarantees a consistency between the indicators resulting from both random experiments. In Section 4, we showcase our new summarization procedure on CDNET and discuss how it affects the ranking between algorithms. Section 5 concludes the paper.

2. PERFORMANCE INDICATORS FOR ONE VIDEO

The performance of a background subtraction algorithm on a video is measured by running it once (even for non-deterministic algorithms), and counting at the pixel level the amounts of true negatives TN ($y = \hat{y} = c^-$), false positives FP ($y = c^-$ and $\hat{y} = c^+$), false negatives FN ($y = c^+$ and $\hat{y} = c^-$), and true positives TP ($y = \hat{y} = c^+$).

The performance indicators are then derived from these amounts. For example, the prior of the positive class is $\pi^+ = \frac{FN+TP}{TN+FP+FN+TP}$, the rate of positive predictions is $\tau^+ = \frac{FP+TP}{TN+FP+FN+TP}$, the error rate is $ER = \frac{FP+FN}{TN+FP+FN+TP}$, the true negative rate (specificity) is $TNR = \frac{TN}{TN+FP}$, the false positive rate is $FPR = \frac{FP}{TN+FP}$, the false negative rate is $FNR = \frac{FN}{FN+TP}$, the true positive rate (recall) is $TPR = R = \frac{TP}{FN+TP}$, the positive predictive value (precision) is $PPV = P = \frac{TP}{FP+TP}$, and the F-score is $F = \frac{2TP}{FP+FN+2TP}$.

The (FPR, TPR) coordinates define the well-known Receiver Operating Characteristic (ROC) space [16, 17]. The

precision P and recall R define the PR evaluation space [13], which is an alternative to the ROC space, as there exists a bijection between these spaces, for a given prior π^+ [11]. In fact, this bijection is just a particular case of the relationships that exist between the various indicators. There are other famous relationships. For example, the F-score is known to be the harmonic mean of the precision P and recall R.

Despite its importance for the interpretation, it is often overlooked that some indicators have a probabilistic meaning [14]. Let us consider the following random experiment.

Random Experiment 1 (for one video) *Draw one pixel at random (all pixels being equally likely) from the video and observe the ground-truth class Y as well as the predicted class \hat{Y} for this pixel. The result of the random experiment is the pair $\Delta = (Y, \hat{Y})$.*

We use capital letters for Y , \hat{Y} , and Δ to emphasize the random nature of these variables. The outcome of this random experiment can be associated with a true negative $tn = (c^-, c^-)$, a false positive $fp = (c^-, c^+)$, a false negative $fn = (c^+, c^-)$, or a true positive $tp = (c^+, c^+)$. The family of *probabilistic indicators* can be defined based on this random experiment:

$$P(\Delta \in \mathcal{A} | \Delta \in \mathcal{B}) \text{ with } \emptyset \subsetneq \mathcal{A} \subsetneq \mathcal{B} \subseteq \{tn, fp, fn, tp\} \quad (1)$$

It includes $\pi^+ = P(\Delta \in \{fn, tp\})$, $\tau^+ = P(\Delta \in \{fp, tp\})$, $ER = P(\Delta \in \{fp, fn\})$, $TNR = P(\Delta = tn | \Delta \in \{tn, fp\})$, $TPR = P(\Delta = tp | \Delta \in \{fn, tp\})$, and $PPV = P(\Delta = tp | \Delta \in \{fp, tp\})$. All the other performance indicators (the F-score, balanced accuracy, etc) can be derived from the probabilistic indicators.

3. SUMMARIZING THE PERFORMANCE INDICATORS FOR SEVERAL VIDEOS

Let us denote a set of videos by \mathbb{V} . We generalize the random experiment defined for one video to several videos as follows.

Random Experiment 2 (for several videos) *First, draw one video V at random in the set \mathbb{V} , following an arbitrarily chosen distribution $P(V)$. Then, draw one pixel at random (all pixels being equally likely) from V and observe the ground-truth class Y and the predicted class \hat{Y} for this pixel. The result of the experiment is the pair $\Delta = (Y, \hat{Y})$.*

As the result of this second random experiment is again a pair of classes, as encountered for a single video (first random experiment), all performance indicators can be defined in the same way. This is important for the interpretability, since it guarantees that all indicators are coherent and that the probabilistic indicators preserve a probabilistic meaning.

The distribution of probabilities $P(V)$ parameterizes the random experiment. It can be arbitrarily chosen to reflect the relative weights given to the videos. For example, one could

use the weights considered in CDNET. An alternative consists in choosing weights proportional to the size of the videos (the number of pixels multiplied by the number of frames). This leads to identical distributions $P(\Delta)$ between the second random experiment and the first one with the fictitiously aggregated video $\bigcup_{v \in \mathbb{V}} v$.

Summarization formulas. We denote by $I(v)$ the value of an indicator I obtained with our first random experiment on a video $v \in \mathbb{V}$, and by $I(\mathbb{V})$ the value of I resulting from our second random experiment applied on the set \mathbb{V} of videos. We claim that the indicators $I(\mathbb{V})$ summarize the performances corresponding to the various videos, because they can be computed based on some $I(v)$, exclusively. To prove it, let us consider a probabilistic indicator $I_{\mathcal{A}|\mathcal{B}}$ defined as $P(\Delta \in \mathcal{A}|\Delta \in \mathcal{B})$, and $I_{\mathcal{B}}$ as $P(\Delta \in \mathcal{B})$. Thanks to our second random experiment, we have

$$I_{\mathcal{A}|\mathcal{B}}(\mathbb{V}) = P(\Delta \in \mathcal{A}|\Delta \in \mathcal{B}) \quad (2)$$

$$= \sum_{v \in \mathbb{V}} P(\Delta \in \mathcal{A}, V = v|\Delta \in \mathcal{B}) \quad (3)$$

$$= \sum_{v \in \mathbb{V}} P(V = v|\Delta \in \mathcal{B}) P(\Delta \in \mathcal{A}|\Delta \in \mathcal{B}, V = v)$$

$$= \sum_{v \in \mathbb{V}} P(V = v|\Delta \in \mathcal{B}) I_{\mathcal{A}|\mathcal{B}}(v), \quad (4)$$

with

$$P(V = v|\Delta \in \mathcal{B}) = \frac{P(V = v) P(\Delta \in \mathcal{B}|V = v)}{P(\Delta \in \mathcal{B})} \quad (5)$$

$$\begin{aligned} &= \frac{P(V = v) P(\Delta \in \mathcal{B}|V = v)}{\sum_{v' \in \mathbb{V}} P(V = v') P(\Delta \in \mathcal{B}|V = v')} \\ &= \frac{P(V = v) I_{\mathcal{B}}(v)}{\sum_{v' \in \mathbb{V}} P(V = v') I_{\mathcal{B}}(v')}. \end{aligned} \quad (6)$$

Thus, the summarized probabilistic indicator $I_{\mathcal{A}|\mathcal{B}}(\mathbb{V})$ is a weighted arithmetic mean of the indicators $\{I_{\mathcal{A}|\mathcal{B}}(v)|v \in \mathbb{V}\}$, the weights being the (normalized) product between the relative importance $P(V = v)$ given to the video v and the corresponding $I_{\mathcal{B}}(v)$:

$$I_{\mathcal{A}|\mathcal{B}}(\mathbb{V}) = \sum_{v \in \mathbb{V}} \frac{P(V = v) I_{\mathcal{B}}(v)}{\sum_{v' \in \mathbb{V}} P(V = v') I_{\mathcal{B}}(v')} I_{\mathcal{A}|\mathcal{B}}(v). \quad (7)$$

In the particular case of an unconditional probabilistic indicator $I_{\mathcal{A}} = I_{\mathcal{A}|\{\text{tn,fp,fn,tp}\}}$, Equation (7) reduces to

$$I_{\mathcal{A}}(\mathbb{V}) = \sum_{v \in \mathbb{V}} P(V = v) I_{\mathcal{A}}(v). \quad (8)$$

Note that we are able to summarize probabilistic indicators that are undefined for some, but not all, videos. The reason is that $I_{\mathcal{A}|\mathcal{B}}(v)$ is undefined only when $I_{\mathcal{B}}(v) = 0$, and we observe that only the product of these two quantities appears

in Equation (7). This product is always well defined since it is an unconditional probability ($I_{\mathcal{B}}(v) I_{\mathcal{A}|\mathcal{B}}(v) = I_{\mathcal{A} \cap \mathcal{B}}(v)$). To emphasize it, we rewrite Equation (7) as

$$\begin{aligned} I_{\mathcal{A}|\mathcal{B}}(\mathbb{V}) &= \sum_{v \in \mathbb{V}} \frac{P(V = v) I_{\mathcal{B}}(v) I_{\mathcal{A}|\mathcal{B}}(v)}{I_{\mathcal{B}}(\mathbb{V})} \\ &= \frac{\sum_{v \in \mathbb{V}} P(V = v) I_{\mathcal{A} \cap \mathcal{B}}(v)}{I_{\mathcal{B}}(\mathbb{V})} = \frac{I_{\mathcal{A} \cap \mathcal{B}}(\mathbb{V})}{I_{\mathcal{B}}(\mathbb{V})}. \end{aligned} \quad (9)$$

Applying our summarization in the ROC (FPR, TPR) and PR ($\mathbf{R} = \text{TPR}$, $\mathbf{P} = \text{PPV}$) spaces. Let us consider the case in which π^+ , τ^+ , FPR, TPR, and PPV are known for each video. The summarized indicators can be computed by Equations (7)-(8), since $\pi^+ = I_{\{\text{fn,tp}\}}$, $\tau^+ = I_{\{\text{fp,tp}\}}$, FPR = $I_{\{\text{fp}\}|\{\text{tn,fp}\}}$, TPR = $I_{\{\text{tp}\}|\{\text{fn,tp}\}}$, and PPV = $I_{\{\text{tp}\}|\{\text{fp,tp}\}}$. This yields:

$$\pi^+(\mathbb{V}) = \sum_{v \in \mathbb{V}} P(V = v) \pi^+(v), \quad (10)$$

$$\tau^+(\mathbb{V}) = \sum_{v \in \mathbb{V}} P(V = v) \tau^+(v), \quad (11)$$

$$\text{FPR}(\mathbb{V}) = \frac{1}{\pi^-(\mathbb{V})} \sum_{v \in \mathbb{V}} P(V = v) \pi^-(v) \text{FPR}(v), \quad (12)$$

$$\text{TPR}(\mathbb{V}) = \frac{1}{\pi^+(\mathbb{V})} \sum_{v \in \mathbb{V}} P(V = v) \pi^+(v) \text{TPR}(v), \quad (13)$$

$$\text{PPV}(\mathbb{V}) = \frac{1}{\tau^+(\mathbb{V})} \sum_{v \in \mathbb{V}} P(V = v) \tau^+(v) \text{PPV}(v). \quad (14)$$

The passage from ROC to PR is given by:

$$\text{PPV}(\mathbb{V}) = \frac{\pi^+(\mathbb{V}) \text{TPR}(\mathbb{V})}{\pi^-(\mathbb{V}) \text{FPR}(\mathbb{V}) + \pi^+(\mathbb{V}) \text{TPR}(\mathbb{V})}. \quad (15)$$

A generic algorithm for the computation of any summarized indicator. Let us assume that the elements of the normalized confusion matrix ($I_{\{\text{tn}\}}$, $I_{\{\text{fp}\}}$, $I_{\{\text{fn}\}}$, and $I_{\{\text{tp}\}}$) can be retrieved for each video. This could be done by computing them based on other indicators. In this case, an easy-to-code algorithm to summarize the performances is the following:

- step 1: arbitrarily weight the videos with $P(V)$,
- step 2: retrieve $I_{\{\text{tn}\}}(v)$, $I_{\{\text{fp}\}}(v)$, $I_{\{\text{fn}\}}(v)$, and $I_{\{\text{tp}\}}(v)$ for each video v ,
- step 3: then compute $I_{\{\text{tn}\}}(\mathbb{V})$, $I_{\{\text{fp}\}}(\mathbb{V})$, $I_{\{\text{fn}\}}(\mathbb{V})$, and $I_{\{\text{tp}\}}(\mathbb{V})$ with Equation (8),
- step 4: and finally derive all the desired summarized indicators based on their relationships with the $I_{\{\text{tn}\}}$, $I_{\{\text{fp}\}}$, $I_{\{\text{fn}\}}$, $I_{\{\text{tp}\}}$ indicators. For example,

$$\text{F}(\mathbb{V}) = \frac{2I_{\{\text{tp}\}}(\mathbb{V})}{I_{\{\text{fp}\}}(\mathbb{V}) + I_{\{\text{fn}\}}(\mathbb{V}) + 2I_{\{\text{tp}\}}(\mathbb{V})}.$$

4. EXPERIMENTS WITH CDNET 2014

We illustrate the summarization with the CDNET 2014 dataset [9] (organized in 11 categories, containing each 4 to 6 videos). We recomputed all the performance indicators for the 36 unsupervised background subtraction algorithms (we could consider all the algorithms as well) whose segmentation maps are given on <http://changedetection.net>, compared to the ground-truth segmentation maps.

Our first experiment aims at determining if the summarized performance is significantly affected by the summarization procedure. Figure 2 compares the results of two summarization procedures, in both the ROC and PR spaces.

1. In the original CDNET procedure, performance indicators are obtained for each category by averaging the indicators measured on each video individually. A final summarized performance is then computed by averaging the performance indicators over the categories.
2. We applied our summarization, with the same weights for each category as in the original setup; thus, for each video, we have $P(V = v) = \frac{1}{11} \times \frac{1}{M}$, where M is the number of videos in the corresponding category.

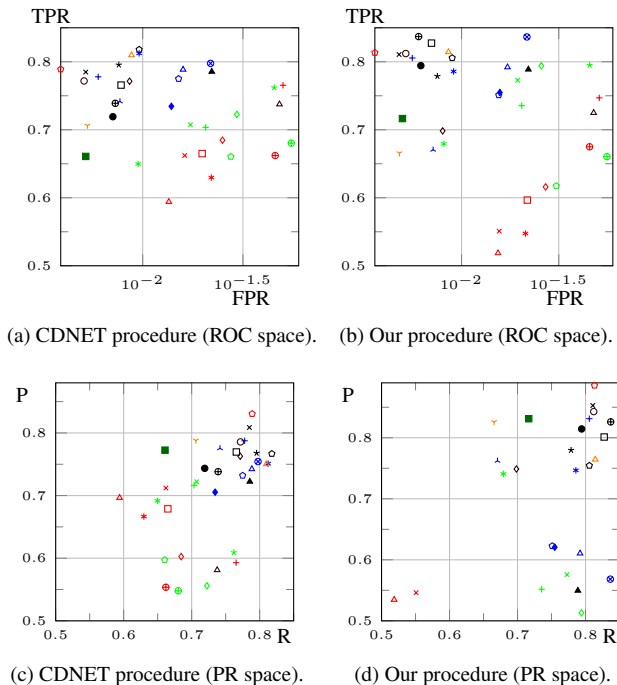


Fig. 2: Summarized performances according to two different procedures, in the cropped ROC (upper row) or PR (lower row) spaces, for 36 classifiers evaluated on the CDNET 2014 dataset. The performances obtained by these procedures differ significantly. Only our summarization procedure preserves the bijection between ROC and PR (see text).

Algorithm	F of [9]	F [ours]
SemanticBGS	0.8098 (1)	0.8479 (1)
IUTIS-5	0.7821 (2)	0.8312 (3)
IUTIS-3	0.7694 (3)	0.8182 (5)
WisenetMD	0.7559 (4)	0.7791 (10)
SharedModel	0.7569 (5)	0.7885 (8)
WeSamBE	0.7491 (6)	0.7792 (9)
SuBSENSE	0.7453 (7)	0.7657 (12)
PAWCS	0.7478 (8)	0.8272 (4)

Table 1: Extract of F-scores (ranks) obtained with two summarization procedures on CDNET.

It can be seen that the results differ significantly (both in ROC and PR), which emphasizes the influence of the summarization procedure for the comparison between algorithms. However, with the original summarization procedure, the intrinsic relationships between indicators are not preserved at the summarized level, while ours preserves them.

Our second experiment aims at determining if the ranking between algorithms depends on the summarization procedure. Because the F-score is often considered as an appropriate ranking indicator for background subtraction, we performed our experiment with it. Table 1 shows the F-scores and ranks obtained according to the same two summarization procedures. We see that the summarization procedure affects the ranking.

5. CONCLUSION

In background subtraction, algorithms are evaluated by applying them on videos and comparing their results to ground-truth references. Performance indicators of individual videos are then combined to derive indicators representative for a set of videos, during a procedure called “summarization”.

We have shown that a summarization procedure based on the arithmetic mean leads to inconsistencies between summarized indicators and complicates the comparison between algorithms. Therefore, based on the definition of a random experiment, we presented a new summarization procedure and formulas that preserve the intrinsic relationships between indicators. These summarized indicators are specific for this random experiment and for the weights given to the various videos. An easy-to-code algorithm for our summarization procedure has been given. Our procedure is illustrated and commented on the CDNET dataset.

As a general conclusion, for background subtraction involving multiple videos, we recommend to always use our summarization procedure, instead of the arithmetic mean, to combine performance indicators calculated separately on each video. By doing so, the formulas that hold between them at the video level also hold at the summarized level.

6. REFERENCES

- [1] T. Bouwmans, F. El Baf, and B. Vachon, “Statistical background modeling for foreground detection: A survey,” in *Handbook of Pattern Recognition and Computer Vision (volume 4)*, chapter 3, pp. 181–199. World Scientific Publishing, January 2010.
- [2] T. Bouwmans, “Traditional and recent approaches in background modeling for foreground detection: An overview,” *Computer Science Review*, vol. 11-12, pp. 31–66, May 2014.
- [3] T. Bouwmans, S. Javed, M. Sultana, and S. Jung, “Deep neural network concepts for background subtraction: A systematic review and comparative evaluation,” *Neural Networks*, vol. 117, pp. 8–66, September 2019.
- [4] S. Elhabian, K. El-Sayed, and S. Ahmed, “Moving object detection in spatial domain using background removal techniques – State-of-art,” *Recent Patents on Computer Science*, vol. 1, pp. 32–54, January 2008.
- [5] P.-M. Jodoin, S. Piérard, Y. Wang, and M. Van Droogenbroeck, “Overview and benchmarking of motion detection methods,” in *Background Modeling and Foreground Detection for Video Surveillance*, chapter 24. Chapman and Hall/CRC, July 2014.
- [6] A. Sobral and A. Vacavant, “A comprehensive review of background subtraction algorithms evaluated with synthetic and real videos,” *Computer Vision and Image Understanding*, vol. 122, pp. 4–21, May 2014.
- [7] N. Vaswani, T. Bouwmans, S. Javed, and P. Narayana-murthy, “Robust subspace learning: Robust PCA, robust subspace tracking, and robust subspace recovery,” *IEEE Signal Processing Magazine*, vol. 35, no. 4, pp. 32–55, July 2018.
- [8] N. Goyette, P.-M. Jodoin, F. Porikli, J. Konrad, and P. Ishwar, “changedetection.net: A new change detection benchmark dataset,” in *IEEE International Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, June 2012.
- [9] N. Goyette, P.-M. Jodoin, F. Porikli, J. Konrad, and P. Ishwar, “A novel video dataset for change detection benchmarking,” *IEEE Transactions on Image Processing*, vol. 23, no. 11, pp. 4663–4679, November 2014.
- [10] C. Cuevas, E. Yanez, and N. Garcia, “Labeled dataset for integral evaluation of moving object detection algorithms: LASIESTA,” *Computer Vision and Image Understanding*, vol. 152, pp. 103–117, November 2016.
- [11] J. Davis and M. Goadrich, “The relationship between precision-recall and ROC curves,” in *International Conference on Machine Learning (ICML)*, Pittsburgh, Pennsylvania, June 2006, pp. 233–240.
- [12] G. Forman and M. Scholz, “Apples-to-apples in cross-validation studies: Pitfalls in classifier performance measurement,” *ACM SIGKDD Explorations Newsletter*, vol. 12, no. 1, pp. 49–57, November 2010.
- [13] K. Boyd, V. Costa, J. Davis, and C. Page, “Unachievable region in precision-recall space and its effect on empirical evaluation,” in *International Conference on Machine Learning (ICML)*, Edinburgh, UK, June-July 2012, pp. 639–646.
- [14] C. Goutte and E. Gaussier, “A probabilistic interpretation of precision, recall and F-score, with implication for evaluation,” in *Advances in Information Retrieval (Proceedings of ECIR)*, Santiago de Compostela, Spain, March 2005, vol. 3408 of *Lecture Notes in Computer Science*, pp. 345–359, Springer.
- [15] J. Bertrand, *Calcul des probabilités*, Gauthier-Villars et fils, 1889.
- [16] P. Flach, “The geometry of ROC space: Understanding machine learning metrics through ROC isometrics,” in *International Conference on Machine Learning (ICML)*, Washington, USA, August 2003, pp. 194–201.
- [17] T. Fawcett, “An introduction to ROC analysis,” *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861–874, June 2006.