Optimal Multiclass Classifier Threshold Estimation with Particle Swarm Optimization for Visual Object Recognition*

Shinko Y. Cheng, Yang Chen, Deepak Khosla, and Kyungnam Kim

HRL Laboratories, LLC
3011 Malibu Canyon Road
Malibu CA 90265
{sycheng,ychen,dkhosla,kkim}@hrl.com

Abstract. We present a novel method to maximize multiclass classifier performance by tuning the thresholds of the constituent pairwise binary classifiers using Particle Swarm Optimization. This post-processing step improves the classification performance in multiclass visual object detection by maximizing the area under the ROC curve or various operating points on the ROC curve. We argue that the precision-recall or confusion matrix commonly used for measuring the performance of multiclass visual object detection algorithms is inadequate to the *Multiclass ROC* when the intent is to apply the recognition algorithm for surveillance where objects remain in view for multiple consecutive frames, and where background instances exists in far greater numbers than target instances. We demonstrate its efficacy on the visual object detection problem with a 4-class classifier. Despite this, the PSO threshold tuning method can be applied to all pairwise multiclass classifiers using any computable performance metric.

1 Introduction

This paper introduces a novel method to maximize multiclass classifier performance by tuning the thresholds of the constituent pairwise binary classifiers using Particle Swarm Optimization (PSO) [1]. This post-processing step improves the classification performance in multiclass visual object detection by maximizing the area under the ROC curve or various operating points on the ROC curve. We argue that the precision-recall or confusion matrix commonly used for measuring the performance of multiclass visual object detection algorithms is insufficient to the *Multiclass ROC* when the intent is to apply the recognition algorithm for surveillance where objects remain in view for multiple consecutive frames, and where background instances exists in far greater

* This work was partially supported by the Defense Advanced Research Projects Agency (government contract no. HR0011-10-C-0033) NeoVision2 program. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressly or implied, of the Defense Advanced Research projects Agency of the U.S. Government.

G. Bebis et al. (Eds.): ISVC 2011, Part II, LNCS 6939, pp. 536–544, 2011.

numbers than target instances. Finally, although we demonstrate its efficacy on the visual object detection problem, this method can be applied to all pairwise multiclass classifiers using any computable performance metric.

A well-established method for constructing a multiclass classifier is by training several binary classifiers, one for each pair of classes among M classes [2,3]. A total of $\binom{M}{2}$ classifiers are trained and applied to a test sample to obtain a prediction. Each pairwise classifier generates a vote for a class. Then, a prediction is made by collecting all the votes and selecting the class with the majority of votes.

The problem with this method is the suboptimal choice of thresholds resulting from considering only two classes of samples at a time in training the constituent binary classifiers. Each pairwise classifier optimally divides the samples between only the two classes in the pair. It is possible to obtain a better mean correct classification performance over all classes by increasing the rate of certain classes with only a small decrease in rate of the other classes. The challenge is in estimating a more optimal threshold for each pairwise classifier. We propose to solve this problem by tuning the thresholds to maximize the recognition performance as a function of the Multiclass ROC, using PSO.

The advantage of PSO for function optimization is that it does not require computing complex, sometimes infeasible derivatives of the objective function. This enlarges the space of objective functions to use. We demonstrate a threshold tuning algorithm utilizing the geometric mean of the areas under the Multiclass ROC of each class, and separately the geometric mean of the *class* true-positive rate at particular false-positive rates. Furthermore, since PSO consists of a single update equation, its implementation is very straightforward.

The remainder of this paper is organized as follows In sec. 1.1, we will review related metrics in literature for evaluating multiclass visual object detection. We will argue that existing metrics paint an incomplete picture largely due to the multiclass aspect of the underlying model. In sec. 2, we describe threshold tuning using PSO with the objective of maximizing Multiclass ROC performance. In sec. 3, we show the efficacy of the optimization, and conclude with a discussion in sec. 4.

1.1 Related Work

A popular metric for multi-class visual object recognition is precision-recall [4]. This metric stems from the problem of image-retrieval, and uses recall to represent the percentage of positive examples successfully retrieved (detected) and precision to measure the percentage of correct instances retrieved. In video surveillance, these measurements are also very important and relevant. However, false-positive rate – which is the percentage of non-target instances erroneously detected as targets – has a distinct advantage in that surveillance analysts can place an expectation on the number of false-positives per unit time by directly scaling the false-positive rate accordingly. However, precision only indirectly

represents false-positives relative to the proportion of the recalled instances. On the other hand, precision does allow the analyst to place an expectation of the number of detected targets to be erroneous.

Variations of the PASCAL VOC metric can be observed in numerous efforts addressing various aspects of the visual object detection problem for surveillance, including Video Analysis and Content Extraction (VACE/CLEAR [5,6]), which pre-dates the PASCAL VOC metric [7], CalTech Pedestrian Detection Benchmark [8] and more recently for events recognition Video Image Retrieval and Analysis Tool (VIRAT) [9]. Despite the level of effort, little attention has been given to performance metrics that captures detection accuracy of a *multiclass* detector. Much of the effort focused on evaluating tracking performance (VACE/CLEAR). Several consider only the two-class problem (VACE/CLEAR, CalTech Pedestrian) or only partially address performance evaluation of the multiclass detector (PASCAL VOC). Specifically, for the VACE/CLEAR and PASCAL VOC metrics, the performance in detecting multiple categories of objects is measured by treating each detector independent of the other, generating one Spatial Frame Detection Accuracy (SFDA) [5] or precision-recall curve [7] per object. This implies that the system's computational complexity and the false-positive rate scales linearly in the number of categories. However, any practical multiclass object recognition system would employ some hierarchical prediction scheme and category confusion resolution mechanism for each detection area; the strengths of the practical system would not be captured by these metrics.

2 Optimal Multi-class Classifiers from Pairwise Binary Classifiers

A multi-class classifier based on pairwise binary classifiers is composed of a pairwise classifier for every combination of pairs among M classes. The set of all pairs of classes can be defined as

$$C^M = \{12, 13, 14, ..., 1M, 23, 24, 25, ..., (M-1)M\} \tag{1}$$

where $|C^M| = \binom{M}{2} = M(M-1)/2$. If a pairwise classifier casts a vote between the ith and jth classes, the classification response for a given sample is defined as $r_{(ij)}(\mathbf{x}), \forall (ij) \in C^M$, and the vector of all pairwise classifier responses is defined as

$$\mathbf{r}(\mathbf{x}) = (r_{12}(\mathbf{x}), ..., r_{(M-1)M}(\mathbf{x}))^\top \tag{2}$$

In the process of voting, each response is compared to a threshold t_{ij} and a vote is cast for one of the two classes. If the response is greater than or equal to t_{ij}, the vote is cast to class i; otherwise, to class j. The class given the most votes among all pairwise classifiers determines to which class the input sample $\mathbf{x}$ is predicted to belong.

2.1 Multiclass ROC

This paper proposes a Multiclass ROC metric for evaluating multi-class classifiers in visual object detection applications. Such classifiers consist of $M-1$ distinct target classes and a single background class. The proposed metric is derived from the 2-class ROC, which explicitly plots the true-positive rate for a given false-positive rate, allowing the algorithm designer to tune the classifier to the desired operating point. Each point on the 2-class ROC is an operating point defined by the value of a threshold applied to the classifier responses. In a similar fashion, the Multiclass ROC has the false-positive and true-positive rate as its two axes; however, there are $M-1$ curves for each of the $M-1$ target classes, and the operating point consists of the set of points on each curve with the same false-positive rate, and the operating point is defined by a set of thresholds applied to the constituent classifiers of the multi-class classifier. Fig. 1 illustrates a conceptual multiclass ROC plot for a multi-class classifier with two target classes and a background class.

In the Multiclass ROC, the curves define functions of the *class* true-positive rate vs. the false-positive rate. If we let $P(i|j)$ be the probability that a sample belonging to class j is predicted as class i, then the class true-positive rate of target class c is defined as

$$P_{\mathrm{d}}(c) = P(c|c) \tag{3}$$

The false-positive rate P_{FA} is the probability of predicting a background sample as a target, and is defined as

$$P_{\mathrm{FA}} = P(\neg B|B) = 1 - P(B|B) = \sum_{c \neq B} P(c|B) \tag{4}$$

Another interpretation of P_{d} and P_{FA} can be derived from the confusion matrix. Each confusion matrix consists of these statistics measured from the predictions made on a (test) dataset with the multiclass classifier and a set of thresholds. The class true-positive rate $P_{\mathrm{d}}(c)$ are the diagonal elements of the confusion matrix for the target classes and the false-positive rate is 1 minus the diagonal element for the background class.

Each operating point consists of a pair of measurements $(P_{\mathrm{d}}, P_{\mathrm{FA}})$. An alternative metric to P_{FA} is false-positives-per-image (FPPI). This metric is also of particular relevance to surveillance, measuring in absolute terms the average

Table 1. Operating points in the Multiclass ROC are derived from values in the Confusion Matrix

	Predicted		
	B	C_1	C_2
B	$P(B\|B) = 1 - P_{\mathrm{FA}}$	$P(C_1\|B)$	$P(C_2\|B)$
Truth C_1	$P(B\|C_1)$	$P(C_1\|C_1) = P_{\mathrm{d}}(C_1)$	$P(C_2\|C_1)$
C_2	$P(B\|C_2)$	$P(C_1\|C_2)$	$P(C_2\|C_2) = P_{\mathrm{d}}(C_2)$

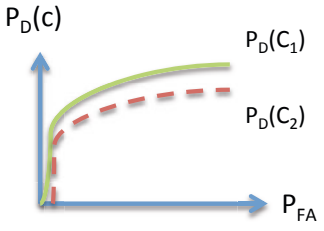


Fig. 1. Concept of the Multiclass ROC curve. Each curve represents the true-positive and false-positive trade-off for a particular class.

number of false positives detected by the classifier per image, which can be rescaled by the imaging frame rate to obtain how frequent in real time a false-positive appears. FPPI can be related to P_{FA} by a normalization factor:

$$P_{\mathrm{FA}} = \frac{\mathrm{FPPI}}{(\text{Total \# background samples/Total \# frames})} \tag{5}$$

To obtain a Multiclass ROC curve, a series of operating points are calculated by using several sets of thresholds for the classifier. In the 2-class case, only a single threshold was required to be varied. In the multi-class case, we vary the thresholds for the of pairwise classifiers that has the potential to vote for the background class. Each step between two threshold values is made equal. More precisely, the pairwise-classifier thresholds are given by $\mathbf{t} = \mathbf{t}_0 + \Delta\mathbf{t}$, where $\mathbf{t}_0$ is the starting threshold value learned with PSO Threshold Tuning described in the sec. 2.2, and swept by the term

$$\Delta\mathbf{t}_{ij} = \begin{cases} t & \text{if } i = B \text{ or } j = B \\ 0 & \text{if } i \neq B \text{ and } j \neq B \end{cases} \tag{6}$$

where $t \in \mathbb{R}$ is varied over the range of classifier responses.

2.2 PSO Threshold Tuning

Because the pairwise classifiers are trained having only considered the samples of two classes and disregarded all of the other classes, subsets of thresholds may be raised or lowered to bias the classifier towards certain classes, e.g. background class in order to minimize false-positives. Furthermore, a bias may raise the true-positive rate for one class at only a slight cost of lowering the true-positive rate for the other classes. As a result, the average true-positive rate among all classes may be raised by optimizing the set of thresholds $\mathbf{t} = (t_{ij}), \forall ij \in C^M$.

Particle Swarm Optimization is used to find the optimal set of thresholds $\mathbf{t}^*$ that maximizes a function of multiclass classification performance. Namely, we solve for

$$\mathbf{t}^* = \arg\max_{\mathbf{t}} f(\mathbf{t}, \mathbf{R}) \tag{7}$$

input : M - Number of classes
$(c1, c2, w)$ - PSO damping factors
N - Number of swarming iterations
P - Number of swarm particles
$\mathbf{R}$ - Set of pairwise classification responses
$f(\mathbf{t}, \mathbf{R})$ - Objective Function
A - Max absolute value classifier responses.
output: $\mathbf{t}^*$ - Optimal Thresholds

```
// Swarm Initialization
```
$\mathsf{dim} \leftarrow M(M-1)/2$;
$g_{\mathsf{conf}} \leftarrow -\infty$;
forall $i \in P$ **do**
  $\mathbf{v}_i \leftarrow \mathbf{0}$;
  $l_{\mathsf{conf},i} \leftarrow -\infty$;
  $\mathbf{x}_i \sim U^{\mathrm{dim}}[-A, A]$;
end
```
// Simulate Swarm
```
for $n \leftarrow 1$ **to** N **do**
  for $i \leftarrow 1$ **to** P `// Update local-best` $\mathbf{l}_i, \forall i$ **do**
    if $l_{\mathsf{conf},i} < x_{\mathsf{conf},i}$ **then**
      $l_{\mathsf{conf},i} \leftarrow x_{\mathsf{conf},i}$;
      $\mathbf{l}_i \leftarrow \mathbf{x}_i$;
    end
  end
  for $i \leftarrow 1$ **to** P `// Update global-best g` **do**
    if $g_{\mathsf{conf}} < l_{\mathsf{conf},i}$ **then**
      $g_{\mathsf{conf}} \leftarrow l_{\mathsf{conf},i}$;
      $\mathbf{g}_i \leftarrow \mathbf{l}_i$;
    end
  end
  for $i \leftarrow 1$ **to** P `// Update particles` $\mathbf{x}_i, \forall i$ **do**
    $r_1 \sim U^{\mathsf{dim}}$;
    $r_2 \sim U^{\mathsf{dim}}$;
    $\mathbf{v}_i \leftarrow w \cdot \mathbf{v}_i + c_1 r_1(\mathbf{l}_i - \mathbf{x}_i) + c_2 r_2(\mathbf{g} - \mathbf{x}_i)$;
    $\mathbf{x}_i \leftarrow \mathbf{x}_i + \mathbf{v}_i$;
    $x_{\mathsf{conf},i} \leftarrow f(\mathbf{x}_i, \mathbf{R})$;
  end
end
return $\mathbf{g}$;

Algorithm 1. PSO Threshold Tuning

where $\mathbf{R} = [\mathbf{r}(\mathbf{x}_1), \mathbf{r}(\mathbf{x}_2), ...\mathbf{r}(\mathbf{x}_N)]$ consists of all pairwise classifier responses from all N samples. PSO was chosen for its simplicity in implementation, flexibility in the kinds of objective functions that can be used, and superior convergence properties compared to similar techniques such as the simplex method or simulated annealing [1]. We define two functions of classification performance based on area under the ROC curve Az (equ. 8) and the true-positive rate P_d (equ. 9).

$$f_{Az}(\mathbf{t}) = \prod_{c=1}^{M} Az(c, \mathbf{t}, \mathbf{R}) \tag{8}$$

$$f_{P_\mathrm{d}}(\mathbf{t}) = \prod_{c=1}^{M} P_d(c, \mathbf{t}, \mathbf{R}) \tag{9}$$

Both $Az(c, \mathbf{t}, \mathbf{R})$ and $P_d(c, \mathbf{t}, \mathbf{R})$ are based on the multiclass ROC metric. Equ. 8 is a function of the *areas* under the receiver-operating-characteristics curve, while equ. 9 is a function of the true-positive rate P_d for a given false-positive rate P_FA. The subscript denotes the class. The objectives using the products rather than the sums of factors would favour a solution that all factors be equally high. The first set of objectives using Az tunes the thresholds to obtain the best classifiers over all operating points, while the second set of objectives using P_d obtains the best classifiers for a given operating point.

The PSO Threshold Tuning routine is given in alg. 1.

3 Experimental Evaluation

An example output from an unoptimized multiclass ROC for a multiclass classifier is shown in fig. 2a. After applying the PSO Threshold Tuning routine with

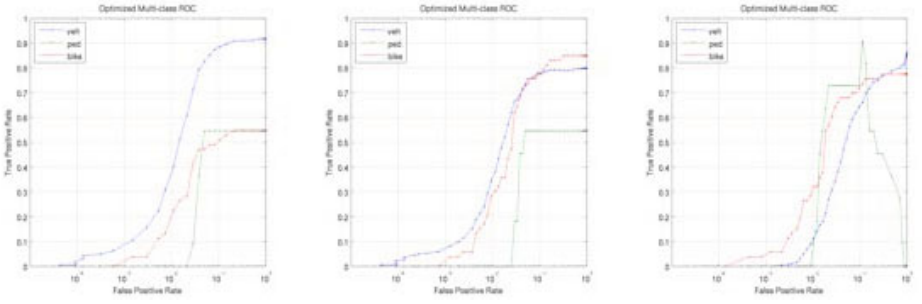
Table 2. Threshold tuning results

(a) Mean Az improved following threshold tuning

Baseline	Az	Optimized Az	Δ
Vehicle	0.8924	0.5904	(0.3020)
Pedestrian	0.5257	0.8635	0.3378
Bike	0.5313	0.7350	0.2037
G.Mean	0.6498	0.7296	0.0798

(b) Mean true-positive rate improved following threshold tuning

Baseline	P_d	Optimized P_d	Δ
Vehicle	0.8806	0.6439	(0.2367)
Pedestrian	0.5455	0.7273	0.1818
Bike	0.5094	0.7170	0.2076
G.Mean	0.6452	0.6961	0.0509



(a) (b) (c)

Fig. 2. Multiclass ROC of a 4-class classifier before and after PSO Threshold-offset tuning. (a) Before tuning. (b) After tuning with objective (2). (c) After tuning with objective (4)

objective function given by equ. 8, the resulting multiclass ROC and class-Az can be seen in fig. 2b and tab. 2a. The geometric mean of areas under the ROC curve Az is used as the objective function for this instance and the tuning procedure raised mean from 0.6294 to 0.7209. As shown in Table 1, the result of tuning lowered the Az for the vehicle class by 0.3020, but it raised the Az for the bike class by 0.2037 and the pedestrian class by 0.3378, which more than the amount lost for the vehicle class.

After applying the PSO Threshold-Offset Tuning routine with the objective given by equ. 9, the resulting multiclass ROC and P_{d} can be seen in fig. 2c and tab. 2a. The geometric mean of the true-positive rates P_{d} at 10% false alarm rate is used as the objective function for this instance. The tuning procedure raised the mean value from 0.1604 to 0.2420. Here, the cost of losing 23% P_{d} in the vehicle is offset by more than 18% increase in P_{d} for both pedestrian and bike classes.

4 Conclusion

We introduced in this paper a novel method to maximize the multiclass classifier performance by tuning the thresholds of the constituent pairwise binary classifiers using Particle Swarm Optimization. We showed that this post-processing step following classifier training improves the performance with respect to the area under the ROC curve or various true-positive rates for a given false-positive rate. In the process, we argued for the use of the Multiclass ROC and demonstrated the efficacy of this approach on a 4-class object classifier.

References

1. Eberhart, R.C., Shi, Y., Kennedy, J.: Swarm Intelligence. Academic Press, London (2001)
2. Duda, R.O., Hart, P.E., Stork, D.G.: Pattern Classification. John Wiley & Sons, Chichester (2001)

3. Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning. Data Mining, Inference, and Prediction, 2nd edn. Springer Series in Statistics (2009)
4. Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The PASCAL visual object classes (VOC) challenge 88, 303–338 (2010)
5. Kasturi, R., Goldgof, D., Soundararajan, P., Manohar, V., Garofolo, J., Bowers, R., Boonstra, M., Korzhova, V., Zhang, J.: Framework for performance evaluation of face, text, and vehicle detection and tracking in video: Data, metrics and protocol. IEEE Transactions on Pattern Analysis and Machine Intelligence 31, 319–336 (2009)
6. Ellis, A., Ferryman, J.M.: PETS2010 and PETS2009 evaluation of results using individual ground truthed single views. In: IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), pp. 135–142 (2010)
7. Everingham, M., Gool, L.V., Williams, C.K.I., Winn, J., Zisserman, A.: The pascal visual object classes (voc) challenge). International Journal of Computer Vision 88, 303–338 (2010)
8. Dollar, P., Wojek, C., Schiele, B., Perona, P.: Pedestrian detection: A benchmark. In: Computer Vision and Pattern Recognition (2009)
9. Oh, S., Perera, A., Cuntoor, N., Chen, C.C., Lee, J.T., Mukherjee, S., Aggarwal, J., Lee, H., Davis, L., Swears, E., Wang, X., Ji, Q., Reddy, K., Shah, M., Vondrick, C., Pirsiavash, H., Ramanan, D., Yuen, J., Torralba, A., Song, B., Fong, A., Roy-Chowdhury, A., Desai, M.: A large-scale benchmark dataset for event recognition in surveillance video. In: IEEE Comptuer Vision and Pattern Recognition (2011)
10. Wu, T.F., Lin, C.J., Weng, R.C.: Probability estimates for multi-class classification by pairwise coupling. Journal of Machine Learning Research 5, 975–1005 (2004)