Evaluation of a binary classifier without Ground-Truth

Grey Kuling[1]

[1]Department of Medical BioPhysics
University of Toronto

May 14, 2017

Medical Biophysics
UNIVERSITY OF TORONTO

Outline

Motivation

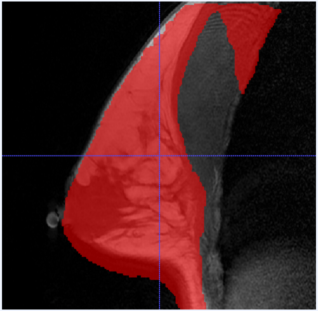


Figure: Decision Tree algorithm Segmentation

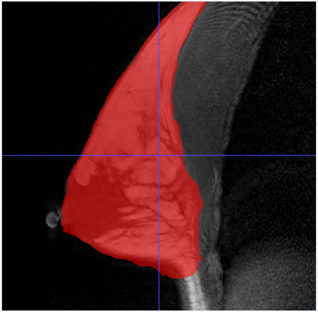


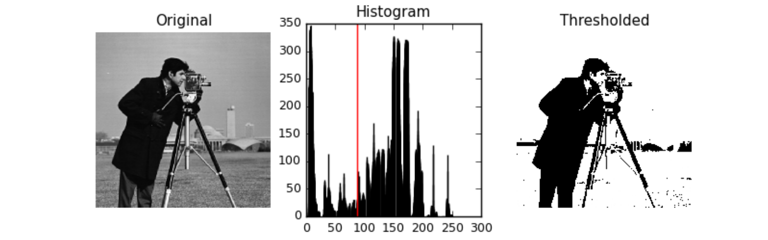
Figure: Manual Segmentation done by presentor

Binary Classifiers

- The task of classifying the elements of a given set Δ into two groups ($C = 0$ or 1) on the basis of a classification rule $S(\cdot)$.
- Example: Thresholding

$$C = S(\delta_i)$$

for all δ_i in Δ



Figure

Binary Classifiers

Figure: Domain of Δ

Binary Classifiers

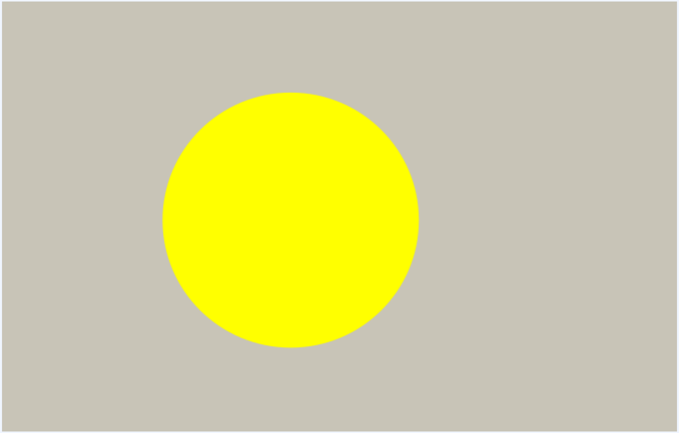


Figure: Ground Truth G on Δ

Binary Classifiers

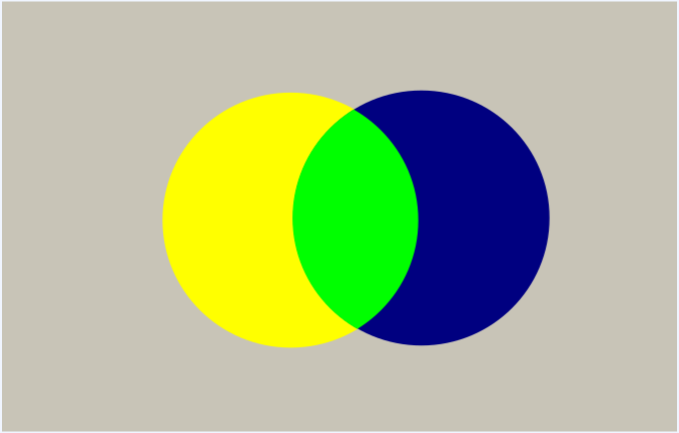


Figure: Ground Truth G with Prediction P over top of Δ

Binary Classifiers

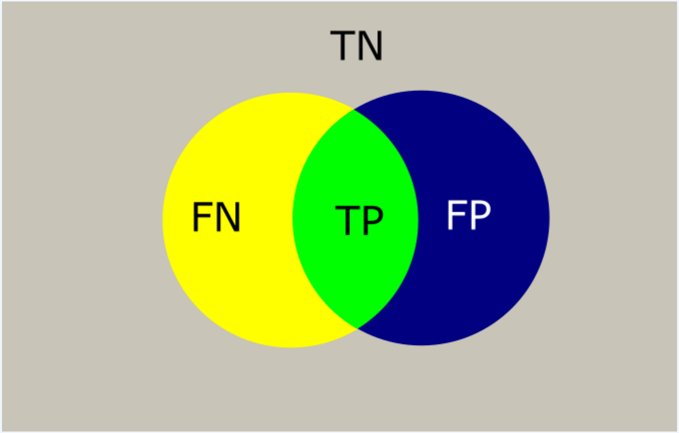


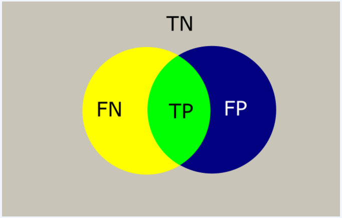
Figure: Confusion Matrix of $S(\cdot)$ on Δ

Precision

- Precision: is the fraction of relevant instances among the retrieved instances. Also referred to as positive predictive value.

$$Pr = \frac{P \cap G}{P} = \frac{TP}{TP + FP}$$

where P: predicted values, and G: ground truth values. While TP: true positives, and FP: false positives.



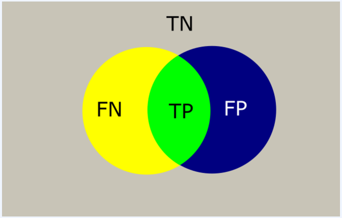
- True or False Test: the amount of correct true answers that you deemed true.

Recall

- Recall: is the fraction of relevant instances that have been retrieved over the total amount of relevant instances.

$$Rc = \frac{P \cap G}{G} = \frac{TP}{TP + FN}$$

where FN: false negatives.



- True or False Test: the amount of correct true answers out of the true facts.

F-Measure

- Precision and Recall are standard metrics expressing the quality of information retrieval methods.
- Also important is the F_β-measure:

$$F_\beta = (1+\beta^2)\frac{PrRc}{\beta^2 Pr + Rc}$$

which is commonly known as the Dice Similarity Coefficient when $\beta = 1$

$$\begin{aligned} F_\beta &= (1+\beta^2)\frac{PrRc}{\beta^2 Pr + Rc} \\ &= \frac{(1+\beta^2)TP}{(1+\beta^2)TP + \beta^2 FN + FP} \\ &= \frac{2TP}{2TP + FP + FN} \end{aligned}$$

Other Metrics

Peak Signal to Noise Ratio (PSNR): is the maximum value between the power of a signal and corrupting noise. The higher this is the better the images match.

$$PSNR = 10 log(\frac{1}{MSE})$$

where

$$MSE = \frac{1}{MN} \sum_{x=1,\ldots,M} \sum_{y=1,\ldots,N} ((G(x,y) - P(x,y))^2$$

Other Metrics

Normalized Cross Correlation: used for comparing multidimensional arrays. The higher this metric the more similar the images are.

$$NCC = \frac{\sum_{x=1,\ldots,M}\sum_{y=1,\ldots,N}(G(x,y)-\overline{G})(P(x,y)-\overline{P})}{\sqrt{\sum_{x=1,\ldots,M}\sum_{y=1,\ldots,N}(G(x,y)-\overline{G})^2\sum_{x=1,\ldots,M}\sum_{y=1,\ldots,N}(P(x,y)-\overline{P})^2}}$$

Other Metrics

Negative Rate Metric (NRM): a numerical equivalent of the relation between mis-classified elements and all other elements in the class. Average of false negative rate and false positive rate. The lower this is the more similar the G and P are.

$$NRM = \frac{FNR + FPR}{2}$$

$$FNR = \frac{FN}{TP + FN}$$

$$FPR = \frac{FP}{TN + FP}$$

What if?

Figure: What if we don't have a reliable G?

Estimators

Some assumptions

1. We are considering a generic system S that given a certain query gives a binary output.

$$S(\delta_i) = 1 \quad \text{or} \quad 0$$

2. Other systems, similar to S exists and their partitioning results are available.

$$S_k(\delta_i) = 1 \quad \text{or} \quad 0$$

Pseudo Precision

- Then in this case each output δ_i will have a probability of being 1

$$P(\delta_i) = \frac{1}{K} \sum_{k=0,1,\ldots,K} S_k(\delta_i)$$

- Important to point out the $k = 0$ is when all $\delta_i = 1$, and $k = K$ is when all $\delta_i = 0$.
- Under the assumption that each δ_i have an equal distribution, we can define precision as the probability that a random document retrieved by a query is relevant.

$$ps\_Pr(S_k) = \frac{\sum_{i=1,\ldots,D} P(\delta_i) S_k(\delta_i)}{\sum_{i=1,\ldots,D} S_k(\delta_i)}$$

Pseudo Recall

- Similarly, Recall can be considered the probability for a random relevant document to be retrieved by the query, and can be found using Bayes' Theorem.

$$
\begin{aligned}
ps\_Rc(S_k) &= P(\text{retrieved by } S_k(\delta_i) | \text{is Relevant}(\delta_i)) \\
&= P(\text{is Relevant}(\delta_i) | \text{retrieved by } S_k(\delta_i)) \frac{P(\text{retrieved by } S_k(\delta_i)}{P(\text{is Relevant}(\delta_i)} \\
&= Pr(S_k) \frac{\frac{1}{D} \sum_{i=1,\dots,D} S_k(\delta_i)}{\frac{1}{D} \sum_{i=1,\dots,D} P(\delta_i)} \\
&= \frac{\sum_{i=1,\dots,D} P(\delta_i) S_k(\delta_i)}{\sum_{i=1,\dots,D} S_k(\delta_i)} \frac{\sum_{i=1,\dots,D} S_k(\delta_i)}{\sum_{i=1,\dots,D} P(\delta_i)} \\
&= \frac{\sum_{i=1,\dots,D} P(\delta_i) S_k(\delta_i)}{\sum_{i=1,\dots,D} P(\delta_i)}
\end{aligned}
$$

Pseudo Precision and Recall

Δ	$P(\delta_i)$	$\mathcal{S}_\top$	$\mathcal{S}_1$	$\mathcal{S}_2$	$\mathcal{S}_3$	$\mathcal{S}_\bot$
δ_1	0.8	1	1	1	1	0
δ_2	0.8	1	1	1	1	0
δ_3	0.4	1	0	1	0	0
δ_4	0.4	1	1	0	0	0
δ_5	0.4	1	1	0	0	0
δ_6	0.4	1	0	0	1	0
δ_7	0.2	1	0	0	0	0
Sum	3.4	7	4	3	3	0
$\sum P\mathcal{S}_k$		3.4	2.4	2	2	0
Pr		0.49	0.6	0.67	**0.67**	∞
Rc		1	**0.71**	0.59	0.59	0

Figure: Example from Lamiroy et al. (2011) of 3 classifiers being compared.

Pseudo Evaluators

pseudo F-measure (DSC):

$$psF_1(S_k) = \frac{2(\sum P(\delta_i) + \sum S_k(\delta_i))}{\sum P(\delta_i) S_k(\delta_i)}$$

pseudo Negative Rate Metric:

$$psNRM = \frac{psFNR + psFPR}{2}$$

$$psFNR = 1 - \frac{\sum P(\delta_i) S_k(\delta_i)}{\sum P(\delta_i)}$$

$$psFNR = \frac{\sum (1 - P(\delta_i)) S_k(\delta_i)}{\sum P(\delta_i)}$$

Pseudo Evaluators

pseudo Normalized Correlation Coefficient

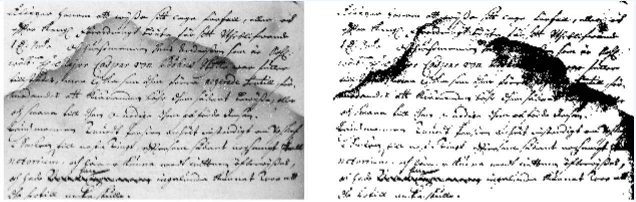
$$psNCC = \frac{\sum_{x=1,\dots,M}\sum_{y=1,\dots,N} S_k(x,y) - \overline{S_k})(P_\delta(x,y) - \overline{P_\delta})}{\sqrt{\sum_{x=1,\dots,M}\sum_{y=1,\dots,N}(S_k(x,y) - \overline{S_k})^2 \sum_{x=1,\dots,M}\sum_{y=1,\dots,N}(P_\delta(x,y) - \overline{P_\delta})^2}}$$

pseudo Peak Signal to Noise Ratio

$$psPSNR = -10log(\frac{1}{MN}\sum_{x=1,\dots,M}\sum_{y=1,\dots,N}(S_k(x,y) - P_\delta(x,y))^2)$$

Evidence from Fedorchuk et al. (2017)

- Digital Image Binarization Contest (DIBCO) Data sets 2009-2013:
- Objective: identify advances in document image binarization by applying evaluation of document image. Collection of images of written words and some of them are corrupted. The goal is binarize them to read the words automatically.



(a) Original Image (b) Binarized (Otsu)

Figure: Example image from DIBCO data set

Evidence from Fedorchuk et al. (2017)

- Used 10 different Thresholding Algorithms: one global (Global Otsu) and nine locally adaptive thresholding algorithms.
- Then calculated the Evaluators compared to the GT, and the pseudo Evaluators and calculated the correlation of conventional evaluators to pseudo Evaluators

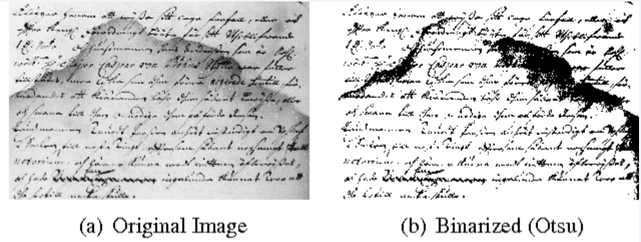


Figure: Example image from DIBCO data set

Evidence from Fedorchuk et al. (2017)

Average correlation coefficient

	FM & ps_FM	PSNR & ps_PSNR	NCC & ps_NCC	NRM & $ps_N RM$
Average	0.845	0.856	0.783	0.373
St. deviation	0.051	0.060	0.234	0.163

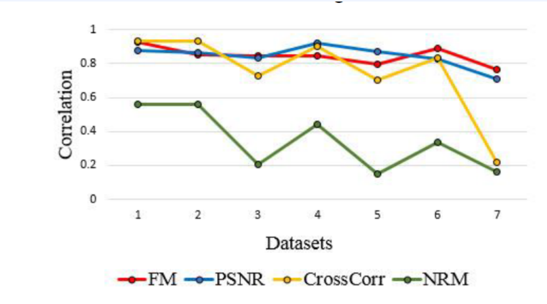


Figure: Correlation Coefficients for different DIBCO data sets

Evidence from Fedorchuk et al. (2017)

Fedorchuk et al. also showed how ell the indicators do with varying amounts of classifiers being used.

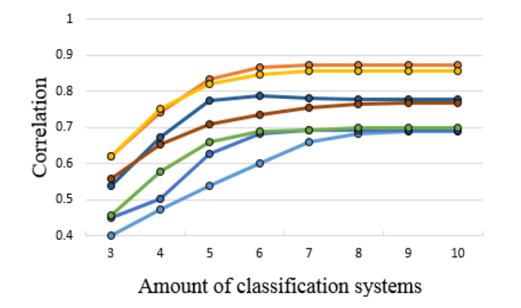


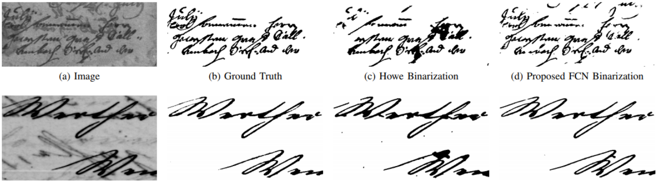
Figure: Correlation Coefficients for different DIBCO data sets

Evidence from Tensmeyer et al. (2017)

- What if we used this pseudo DSC to optimize a NN for segmentaiton. Tensmeyer et al. Gave this a shot.
- Created a 5 layered Fully connected convolutional neural network and used different loss functions (p-FM, FM, p-FM+FM, and Cross entropy) on two different data sets similar to the DIBCO sets.
- He tried this because the new metric for the DIBCO dataset challenges is now the pseudo-FM.

Evidence from Tensmeyer et al. (2017)

- Performances came out quite interestingly.



(a) Image (b) Ground Truth (c) Howe Binarization (d) Proposed FCN Binarization

(e) Image (f) Ground Truth (g) Howe Binarization (h) Proposed FCN Binarization

Figure 2. Qualitative comparison of proposed ensemble of FCNs with state-of-the-art Howe Binarizataion [9]. Images contain significant bleed through noise and come from the H-DIBCO 2016 test data.

Dataset	Loss	Metrics			
		P-FM	FM	DRD	PSNR
HDIBCO 2016	P-FM	**94.09 (94.67)**	86.66 (87.06)	4.62 (4.38)	17.73 (17.86)
	FM	92,90 (93.23)	89.93 (90.30)	3.69 (3.51)	**18.73 (18.90)**
	P-FM + FM	93.22 (93.76)	89.01 (89.52)	4.01 (3.76)	18.48 (18.67)
	Cross-Entropy	92.59 (92.94)	**90.20 (90.56)**	**3.62 (3.45)**	18.68 (18.84)
PLM	P-FM	68.23 (68.55)	66.93 (67.20)	9.24 (9.10)	14.79 (14.83)
	FM	67.40 (67.74)	**68.38 (68.69)**	9.86 (9.68)	14.59 (14.64)
	P-FM + FM	**68.54 (68.96)**	68.27 (68.63)	**9.12 (8.94)**	**14.81 (14.87)**
	Cross-Entropy	66.41 (66.77)	65.38 (65.68)	9.95 (9.78)	14.58 (14.63)

Table I
AVERAGE PERFORMANCE OF 5 FCNS ON H-DIBCO 2016 AND PLM DATASETS FOR VARIOUS LOSS FUNCTIONS. NUMBERS IN PARENTHESIS INDICATE ENSEMBLE PERFORMANCE.

Figure: Results of the FCN from Tensmeyer et al.

Summary

- Reviewed classic evaluators of binary classifiers
- Went through the proofs from Lamiroy et al. for calculating pseudo evaluators of different binary classifiers
- Looked at recent research work giving experimental evidence to the validity of these pseudo evaluators.

Thank you for your time! I'd be happy to answer any questions I can!