Margin Noise Evaluation Metrics

Soumyadeep Dey and Jayanta Mukherjee and Shamik Sural

August 10, 2017

Abstract

The metrics for the margin noise evaluation are defined in this paper.

1 Metrics Used for Evaluation

For evaluation of margin noise removal techniques, seven different metrics are used. They are briefly presented below.

1.1 Hamming Distance ($\mathcal{HD}$)

Hamming distance ($\mathcal{HD}$) is computed as in [1]:

$$\mathcal{HD} = \frac{I_g^t \oplus I_c}{n_f^g} \tag{1}$$

Here, I_g^t is the text part of the groundtruth image, I_c is the cleaned image, $\oplus$ represent exclusive OR operator, and n_f^g is the number of foreground pixels present in the groundtruth image. Smaller the value of $\mathcal{HD}$ better is the quality of the result of processing. This metric computes the proximity of the cleaned image to the textual content of the input document.

1.2 Noise Ratio ($\mathcal{NR}$)

Amount of margin noise present in a page is measured using *noise ratio*. This is defined in [2]:

$$\mathcal{NR} = \frac{n_m^c}{n_t^c} \tag{2}$$

Here, n_m^c and n_t^c, respectively, represent the number of noisy pixels and the number of text pixels present in the cleaned image I_c. This measure helps to identify the amount of noise present in the cleaned image with respect to the actual content of the image. Smaller *noise ratio* indicates better efficiency of an algorithm to remove noise from an image without any penalty for the removal of the original content of the image.

1.3 Page Content Removal ($\mathcal{PR}$)

Amount of page content removed by an algorithm is quantified using *page content removal* measure. This measure is defined as [2]:

$$\mathcal{PR} = \frac{n_t^g - n_t^c}{n_t^g} \tag{3}$$

In Eq. 3, n_t^g and n_t^c represent number of text pixels (original page content) present in the groundtruth image and number of those text pixels also present in the cleaned image, respectively. For a good technique this value should be small.

1.4 Metrics based on Confusion Matrix

As we need to perform detection of noisy texts and other elements around the margin of the main content of a page for removing them, margin noise removal algorithm is also considered as a bi-classification problem. Here, text and graphics part of a document are considered as positive classes, and noise (margin noise) part of the document is considered as negative class. Based on this characterization, a confusion matrix is computed for an image. A sample confusion matrix is shown in Table 1, where tp, fp, tn, and fn represent number of true positives, false positives, true negatives, and false negatives, respectively. Using this confusion matrix, four statistical metrics are defined to quantify the performance of margin noise removal algorithm.

Table 1: Format of a confusion matrix

		output	
		text	noise
input	text	tp	fn
	noise	fp	tn

$FScore_1^{pr}$ This measure is a statistical measure to quantify how positive class of the data is related to the data which are labeled positive by the algorithm. It is the harmonic mean of precision and recall and is defined as [3]:

$$FScore_1^{pr} = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \tag{4}$$

Here, precision $= \frac{tp}{tp+fp}$ and recall $= \frac{tp}{tp+fn}$. This metric measures the ability of an algorithm to classify the positive class.

$FScore_1^{ns}$ This metric is computed to quantify how negative class of the data is related to the data labeled as negative by the concerned algorithm. $FScore_1^{ns}$ is the harmonic mean of *negative predictive value* and specificity. It is computed using Eq. 5, where negative predictive value $= \frac{tn}{tn+fn}$ and specificity $= \frac{tn}{tn+fp}$.

$$FScore_1^{ns} = 2 \times \frac{\text{negative predictive value} \times \text{specificity}}{\text{negative predictive value} + \text{specificity}} \tag{5}$$

The ability of a algorithm to classify negative class is measured using this metric.

Accuracy (*Acc*) Overall performance of an algorithm is measured using this metric and it is defined as:

$$\text{Acc} = \frac{tp + tn}{tp + fp + tn + fn} \tag{6}$$

It may be noted that this metric is biased to the performance of an algorithm with respect to its ability to classify the class with higher number of elements.

Balanced Accuracy (*BalAcc*) Balanced accuracy is a measure to quantify the algorithm's ability to avoid false classification. It is measured by computing average value of the sensitivity (recall) and specificity, and it is defined as [3]:

$$BalAcc = \frac{\text{recall} + \text{specificity}}{2} \tag{7}$$

This measure overcome the biassness of the metric defined in Eq. 6, since equal priority is given to both positive and negative classes.

References

[1] F. Shafait and T. M. Breuel. A simple and effective approach for border noise removal from document images. In *2009 IEEE 13th International Multitopic Conference*, pages 1–5, Dec 2009.

[2] F. Shafait and T. M. Breuel. The effect of border noise on the performance of projection-based page segmentation methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(4):846 – 851, April 2011.

[3] M. Sokolova and G. Lapalme. A systematic analysis of performance measures for classification tasks. *Information Processing Management*, 45(4):427–437, jul 2009.