Multilabel Evaluation Measure Confusion

Rainer Kelz (rainer.kelz@jku.at)

January 2, 2017

1 Multi-F_0 Evaluation Measures

What follows are some of the different definitions for precision, recall, f-measure and accuracy ($\mathcal{P}, \mathcal{R}, \mathcal{F}, \mathcal{A}$), in the multilabel context, according to the scikit-learn package. We use the same naming scheme. The evaluations differ in how errors are treated.

$$\mathcal{P}_{micro} = \frac{\sum_{t=1}^{T} TP[t]}{\sum_{t=1}^{T} TP[t] + FP[t]} \qquad \mathcal{P}_{samples} = \frac{1}{T}\sum_{t=1}^{T} \frac{TP[t]}{TP[t] + FP[t]}$$

$$\mathcal{R}_{micro} = \frac{\sum_{t=1}^{T} TP[t]}{\sum_{t=1}^{T} TP[t] + FN[t]} \qquad \mathcal{R}_{samples} = \frac{1}{T}\sum_{t=1}^{T} \frac{TP[t]}{TP[t] + FN[t]}$$

$$\mathcal{F}_{micro} = \frac{2 \cdot \mathcal{P}_{micro} \cdot \mathcal{R}_{micro}}{\mathcal{P}_{micro} + \mathcal{R}_{micro}} \qquad \mathcal{F}_{samples} = \frac{2 \cdot \mathcal{P}_{samples} \cdot \mathcal{R}_{samples}}{\mathcal{P}_{samples} + \mathcal{R}_{samples}}$$

$$\mathcal{A}_{micro} = \frac{\sum_{t=1}^{T} TP[t]}{\sum_{t=1}^{T} TP[t] + FP[t] + FN[t]} \qquad \mathcal{A}_{samples} = \frac{1}{T}\sum_{t=1}^{T} \frac{TP[t]}{TP[t] + FP[t] + FN[t]}$$

The "micro" averaging scheme is the one used for the MIREX challenge, and is defined in [1]. The main difference between "micro" and "samples" averaging schemes is the treatment of errors. In the "micro" scheme, every error is treated equally (with the same weight), whereas in the "samples" scheme, the weighing depends on the actual amount of errors per frame. Consider the example in figure 1:

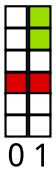

Figure 1: Example drawing of two frames, depicting **true positives** in green and **false positives** in red color.

"samples" scheme	frame 0	frame 1	$\mathcal{P}_{samples}$
	0	2/3	(0 + 2/3) / 2 = 0.3333...

"micro" scheme	frame 0	frame 1	$\mathcal{P}_{micro}$
	-	-	2 / 4 = 0.5

For the "samples" scheme, the false positive in frame 0 is weighed much higher than the false positive in frame 1, whereas for the "micro" scheme, both false positives are weighted equally.

2 Multi-F_0 Evaluation Shenanigans

Following a "chain of unfortunate events", this is the mess in all its ugliness:

[4] The formulas reported in the evaluation section are actually for "samples" averaging, and missing the scaling factor in front. The evaluation actually uses "micro" averaging and compares with results in [6], which are **very likely** using the "micro" averaged definitions as well.

[6] The formulas reported in the evaluation section are actually for "samples" averaging, and missing the scaling factor in front. They compare with "micro" averaged results, and due to the performance numbers reported, its again **very likely** they **actually** use the "micro" averaging scheme.

[1] Defines $\mathcal{P}, \mathcal{R}, \mathcal{F}, \mathcal{A}$ with "micro" averaging, cited in [6] as the source of the evaluation measures (which indicates the formulas in [6] are simply the wrong ones, as in [4], which blindly propagated this error).

[7] Defines $\mathcal{P}, \mathcal{R}, \mathcal{F}, \mathcal{A}$ with "micro" averaging, cited and compared against in [6]

[2] Only accuracy used with "micro" averaging, same as in [3], uses "micro" averaging definition in [3], cited and compared against in [6].

[5] Only accuracy used with "micro" averaging, uses "micro" averaging definition from [3], cited in [6].

[3] Defines/uses accuracy only with "micro" averaging.

References

[1] Mert Bay, Andreas F Ehmann, and J Stephen Downie. Evaluation of multiple-f0 estimation and tracking systems. In *ISMIR*, pages 315–320, 2009.

[2] Emmanouil Benetos and Simon Dixon. A shift-invariant latent variable model for automatic music transcription. *Computer Music Journal*, 36(4):81–94, 2012.

[3] Simon Dixon. On the computer recognition of solo piano music. In *Australasian Computer Music Conference*, pages 31–37, 2000.

[4] Rainer Kelz, Matthias Dorfer, Filip Korzeniowski, Sebastian Böck, Andreas Arzt, and Gerhard Widmer. On the potential of simple framewise approaches to piano transcription. In *ISMIR*, pages 475–482, 2016.

[5] Graham E Poliner and Daniel PW Ellis. A discriminative model for polyphonic piano transcription. *EURASIP Journal on Applied Signal Processing*, 2007(1):154–154, 2007.

[6] Siddharth Sigtia, Emmanouil Benetos, and Simon Dixon. An end-to-end neural network for polyphonic piano music transcription. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(5):927–939, 2016.

[7] Emmanuel Vincent, Nancy Bertin, and Roland Badeau. Adaptive harmonic spectral decomposition for multiple pitch estimation. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(3):528–537, 2010.