# Simplifying the Visualization of Confusion Matrix

Emma Beauxis-Aussalet [a] Lynda Hardman [a]

[a] *CWI - Information Access Group - Science Park 123 - Amsterdam*

**Abstract**

Supervised Machine Learning techniques can automatically extract information from a variety of multimedia sources, e.g., image, text, sound, video. But it produces imperfect results since the multimedia content can be misinterpreted. Errors are commonly measured using confusion matrices, encoding type I and II errors for each class. Non-expert users encounter difficulties in understanding and using confusion matrices. They need to be read both column- and row-wise, which is tedious and error prone, and their technical concepts need explanations. Further, the visualizations commonly use of complex metrics, e.g., Precision/Recall, F1 scores. These can be overwhelming and misleading for non-experts since they may be inappropriate for specific use cases. For instance, type II errors (False Negative) are critical for medical diagnosis while type I errors (False Positive) are more tolerated. In the case of optical sorting of manufactured products (defect detection), the sensitivity to errors can be the opposite. We propose a novel visualization design that address the needs of non-experts users. Our visualization is intended to be easier to understand, and to minimize the risk of misinterpretation, and so for all kind of use cases. Future work will evaluate our design with both experts and non-experts, and compare its effectiveness with that of traditional ROC and Precision/Recall curves.

## 1 Use cases

Confusion matrices are the major mean to evaluate errors in classification problems. They encode the complete specification of misclassifications: the numbers of misclassified items for each pair {original class in which items should be classified, incorrect class in which items are erroneously classified}. Confusion matrices are used for: i) inspecting errors for each class; ii) tuning software parameters such as detection thresholds; iii) comparing software versions. The selection of software version, or parameter settings, basically rely on the tolerance to *Type I or II* errors. The sensitivity to either error type depends on application domains. In some domains, type I are critical while type II are more tolerated: e.g., fraud detection involving automatic suspension of services (bank, mail, social media), biometric identification, recommendation, optical sorting (*Case A*). In other domains, type II are critical while type I are more tolerated: medical diagnosis, threat detection (*Case B*). Others are sensitive to both error types: character recognition, monitoring of population dynamics in ecology (*Case C*).

Analyzing confusion matrices is complicated since FN for one class are FP for another. Users need to inspect the matrix both column-wise and row-wise, and they can forget cell values, or may read only columns or rows. Confusion matrices are usually synthesized by cumulating misclassifications into FP, FN, TP and TN for each class. But users can no longer distinguish which classes are likely to be confused with another. This is an issue in domains analyzing trends over time (e.g., population dynamics in ecology). E.g., an important increase of one class imply an increase of its FN, and can induce a deceiving increase of other classes. Confusion matrices are synthesized further by deriving advanced metrics, e.g., TP rate, FP rates, F1 scores. Non-experts may not know which metrics suit their use case, or misinterpret them. E.g., high TN may conceal critical errors by yielding low *FP Rate* and high *Accuracy*. *Precision* does not convey the errors critical for *Case A*, nor *Recall* and *FP Rate* for *Case B*, nor *Accuracy* and *F1 score* convey the errors critical for neither *Case A and B*. For *Case C*, using only one metrics amongst *Precision, Recall and FP Rate* does not convey sufficient information. ROC and Precision/Recall curves, increase the risk to overwhelm and confuse users. Non-experts may not identify the point offering the best tradeoff between type I and II errors. Further, classifiers' errors can appear identical in one type of curve, or almost identical, while another view would reveal further differences (Fig. 1).

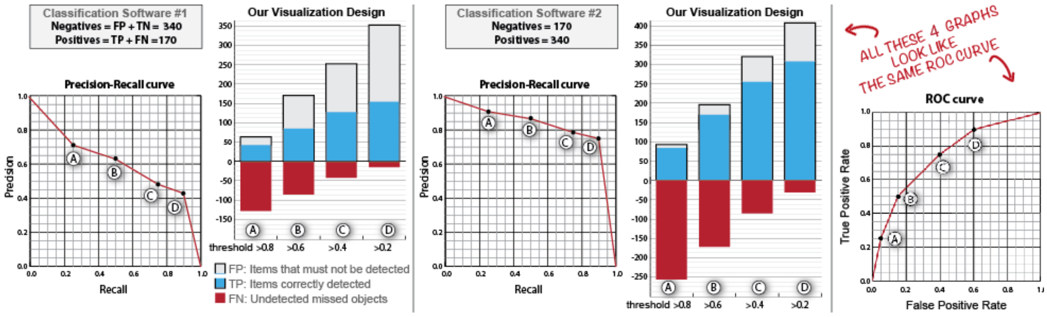

Figure 1: Alternative visualizations: our design, and equivalent ROC and Precision/Recall curves.

## 2 Visualization Design

To ease non-experts' interpretation, we display only TP, FN and FP (Fig. 1). TN are omitted since they are potentially uninteresting (e.g., not contained in end-results) and misleading (e.g., high TN yield high *Accuracy*). We primarily show raw numbers of errors (e.g., in Fig. 1) which are more tangible, without rates and advanced metrics. It preserves the numbers of items in the ground-truth, hidden in ROC and Precision/Recall curves. It displays both type I and II errors, without attempting to show rates relatively to case-dependent frames of reference. Hence it suits most use cases.

Classes can have heterogeneous numbers of ground-truth items, thus being difficult to compare. Hence our design in Fig. 2 reports errors proportionally to the TP of each class. It details inter-classes confusions by indicating the 2 classes producing most FP and FN, and groups errors for all other classes. It indicates potential biases in end-results (e.g., high confusion between classes yielding correlated but unrepresentative trends in end-results). Such information are lost with the usual metrics, since FN and FP are cumulated over classes.

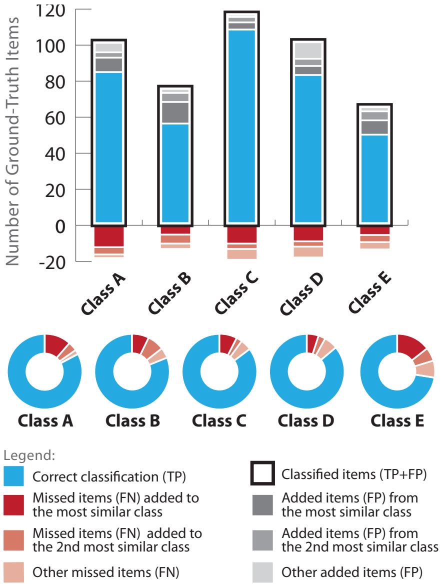

Figure 2: Visualization of inter-classes confusions.

Our design is of interest for Machine Learning researchers and practitioners dealing with the acceptance of their software by end-users. Novel applications of Machine Learning techniques face trust issues: users need to assess their reliability, while evaluations provided by experts are perceived as abstruse, thus impairing further user trust. Our visualization provides a solution for fully conveying the performance of Machine learning software, while minimizing user cognitive effort and remaining intuitive [1, 2]. It can also be used by Machine Learning experts themselves.

## References

[1] E. Beauxis-Aussalet, E. Arslanova, L. Hardman, and J. Van Ossenbruggen. A case study of trust issues in scientific video collections. In *Proceedings of the 2nd ACM international workshop on Multimedia Analysis for Ecological Data*, 2013.

[2] E. Beauxis-Aussalet and L. Hardman. Visualization of confusion matrix for non-expert users. In *Proceedings of the IEEE Symposium on Information Visualization (IEEE InfoVis), in press*, 2014.