An F-Measure for Evaluation of Unsupervised Clustering with Non-Determined Number of Clusters

Ricard Marxer* and Hendrik Purwins
Universitat Pompeu Fabra, Ocata 1, 08001 Barcelona, Spain
Version 0.2

August 22, 2008

Abstract

In unsupervised learning, such as clustering, the problem occurs how to evaluate the results. In particular, neither the number of clusters nor the mapping between eventually known reference classes, e.g. generated from annotations, and the clusters are known. In this report, a method is suggested that adapts the F-measure for supervised classification to the unsupervised case.

The task is to group items into k cluster, without knowing k beforehand. If we have labels to these items (not used in the actual clustering), we can evaluate the unsupervised clustering processes. We adapt a measure introduced by [2] that differs from traditional Receiver Operating Characteristics (ROC). ROC measures have often been used in onset detection evaluation and non-incremental clustering. However, in an unsupervised clustering setting, the mapping between the reference classes and the estimated clusters is unknown.

Analogously to the confusion matrix in the evaluation of a classification tasks with known mapping, we introduce a mapping matrix that is first constructed by using the onset matching technique presented in [1] adapted to multiple classes. The false positives are treated as matches to an extra empty class in the mapping matrix. Similarly, the false negatives are assigned to an empty cluster. Empty classes and clusters are treated in the same as the other classes and clusters with the exception that their precision and recall do not contribute to the overall precision and recall. Then, the measure considers several hypotheses of class-to-cluster assignments and integrates them by taking averages of their precision and recall measures weighted by each class frequency of occurrence.

Let us consider C ground truth classes and K clusters. Be $n_{c,k}$ the number of co-occurrences of class c and cluster k, n_c the total number of occurrences of class c. We

[*] Corresponding author. Email: rmarxer@iua.upf.edu

can calculate the recall of non-c elements within cluster k as $\frac{\sum_{1\leq i\leq C, i\neq c} n_{i,k}}{\sum_{1\leq i\leq C, i\neq c} n_i}$. Then we can express the precision as

$$P(c,k) = \begin{cases} 1 - \frac{\sum_{1\leq i\leq C, i\neq c} n_{i,k}}{\sum_{1\leq i\leq C, i\neq c} n_i} & \text{if } C > 1 \\ 1 & \text{otherwise.} \end{cases} \tag{1}$$

In the second case, there is only one class, yielding always perfect precision. We define the recall as

$$R(c,k) = \begin{cases} \frac{n_{c,k}-1}{n_c-1} & \text{if } n_c > 1 \\ 1 & \text{otherwise.} \end{cases} \tag{2}$$

A cluster k with no instances of class c yields a recall $R(c,k) = 1$. If cluster k has at least one instance of class c, then all the rest $(n_{c,k} - 1)$ should be also from class c.

The pairs of precision and recall of each cluster are integrated to achieve precision and recall measures per class. This is done by summation across the clusters, weighted by the number of occurrences of the class in each of the clusters:

$$P(c) = \frac{\sum_{1\leq k\leq K} n_{c,k} P(c,k)}{n_c} \tag{3}$$

$$R(c) = \frac{\sum_{1\leq k\leq K} n_{c,k} R(c,k)}{n_c} \tag{4}$$

The total precision and recall measures are the weighted sums of the per-class measures:

$$P = \frac{\sum_{1\leq c\leq C} n_c P(c)}{\sum_{1\leq c\leq C} n_c} \tag{5}$$

$$R = \frac{\sum_{1\leq c\leq C} n_c R_{IC}(c)}{\sum_{1\leq c\leq C} n_c} \tag{6}$$

In contrast to using the straight forward definitions of $P(c,k) = \frac{n_{c,k}}{\sum_{1\leq i\leq C} n_{i,k}}$ and $R(c,k) = \frac{n_{c,k}}{n_c}$, an F-measure following the above definition can reach 0 for trivial solutions.

We can use P and R to evaluate the transcription and prediction accuracy, e.g. in unsupervised transcription of music.

This work is partially funded by the *EmCAP* project (European Commission FP6-IST, contract 013123).

References

[1] Paul Brossier. *Automatic Annotation of Musical Audio for Interactive Applications*. PhD thesis, Centre for Digital Music, Queen Mary University of London, London, UK, sep 2006.

[2] Ricard Marxer, Piotr Holonowicz, Hendrik Purwins, and Amaury Hazan. Dynamical hierarchical self-organization of harmonic, motivic, and pitch categories. In *Music, Brain and Cognition. Part 2: Models of Sound and Cognition, held at NIPS)*. Vancouver, Canada, 2007.