

Evaluating the Clustering Results

In order to evaluate the clustering results got from different clustering methods / distance measures, we compare the results to the “ground-truth”. Fortunately, all the datasets we test on have class labels. And they can serve as the “ground-truth”.

Given the clusterings $C = C_1 \dots C_K$ (say the “ground-truth”) and the clusterings $C' = C'_1 \dots C'_K$ (say the results of a particular clustering approach), we can compute the similarity between them using the following formula:

$$Sim(C_i, C'_j) = 2 \frac{|C_i \cap C'_j|}{|C_i| + |C'_j|}$$

and

$$Sim(C, C') = \left(\sum_i \max_j Sim(C_i, C'_j) \right) / K$$

Note that this similarity measure will return 0 if the two clusterings are completely different and 1 if they are identical. The measure was proposed by Gavrilov et al. in [1] and has been used for comparing different clusterings in [2]. Note that this measure is not symmetric. We always use the “ground-truth” clusterings as the first parameter in our implementation.

- [1] M. Gavrilov, D. Anguelov, P. Indyk, and R. Motwani. Mining the stock market: Which measure is best? In *Proc. of KDD'00*, 487-496, 2000.
- [2] B. Larsen, C. Aone. Fast and effective text mining using linear-time document clustering. In *Proc. of KDD'99*, 16-22, 1999.