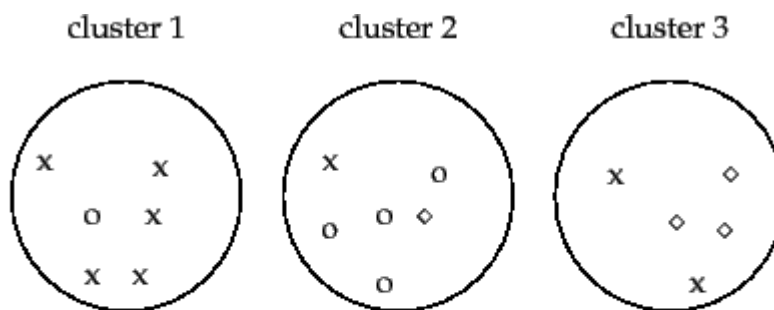# Evaluation of clustering

Typical objective functions in clustering formalize the goal of attaining high intra-cluster similarity (documents within a cluster are similar) and low inter-cluster similarity (documents from different clusters are dissimilar). This is an *internal criterion* for the quality of a clustering. But good scores on an internal criterion do not necessarily translate into good effectiveness in an application. An alternative to internal criteria is direct evaluation in the application of interest. For search result clustering, we may want to measure the time it takes users to find an answer with different clustering algorithms. This is the most direct evaluation, but it is expensive, especially if large user studies are necessary.

As a surrogate for user judgments, we can use a set of classes in an evaluation benchmark or gold standard (see Section 8.5 , page 8.5 , and Section 13.6 , page 13.6). The gold standard is ideally produced by human judges with a good level of inter-judge agreement (see Chapter 8 , page 8.1). We can then compute an *external criterion* that evaluates how well the clustering matches the gold standard classes. For example, we may want to say that the optimal clustering of the search results for jaguar in Figure 16.2 consists of three classes corresponding to the three senses car, animal, and operating system. In this type of evaluation, we only use the partition provided by the gold standard, not the class labels.

This section introduces four external criteria of clustering quality. *Purity* is a simple and transparent evaluation measure. *Normalized mutual information* can be information-theoretically interpreted. The *Rand index* penalizes both false positive and false negative decisions during clustering. The *F measure* in addition supports differential weighting of these two types of errors.

▶ **Figure 16.1**   Purity as an external evaluation criterion for cluster quality.   Majority class and number of members of the majority class for the three clusters are: x, 5 (cluster 1); o, 4 (cluster 2); and ◇, 3 (cluster 3). Purity is $(1/17) \times (5+4+3) \approx 0.71$.

To compute *purity* , each cluster is assigned to the class which is most frequent in the cluster, and then the accuracy of this assignment is measured by counting the number of correctly assigned documents and dividing by N. Formally:

$$\text{purity}(\Omega, \mathbb{C}) = \frac{1}{N} \sum_k \max_j |\omega_k \cap c_j|$$

where $\Omega = \{\omega_1, \omega_2, \ldots, \omega_K\}$ is the set of clusters and $\mathbb{C} = \{c_1, c_2, \ldots, c_J\}$ is the set of classes. We interpret ω_k as the set of documents in ω_k and c_j as the set of documents in c_j in Equation 182.

We present an example of how to compute purity in Figure 16.4 . ☑ Bad clusterings have purity values close to 0, a perfect clustering has a purity of 1 . Purity is compared with the other three measures discussed in this chapter in Table 16.2 .

Table 16.2: The four external evaluation measures applied to the clustering in Figure 16.4

.

	purity	NMI	RI	F_5
lower bound	0.0	0.0	0.0	0.0
maximum	1	1	1	1
value for Figure 16.4	0.71	0.36	0.68	0.46

High purity is easy to achieve when the number of clusters is large - in particular, purity is 1 if each document gets its own cluster. Thus, we cannot use purity to trade off the quality of the clustering against the number of clusters.

A measure that allows us to make this tradeoff is *normalized mutual information* or *NMI* :

$$\text{NMI}(\Omega, \mathbb{C}) = \frac{I(\Omega; \mathbb{C})}{[H(\Omega) + H(\mathbb{C})]/2} \tag{183}$$

I is mutual information (cf. Chapter 13 , page 13.5.1):

$$I(\Omega; \mathbb{C}) = \sum_k \sum_j P(\omega_k \cap c_j) \log \frac{P(\omega_k \cap c_j)}{P(\omega_k) P(c_j)} \tag{184}$$

$$= \sum_k \sum_j \frac{|\omega_k \cap c_j|}{N} \log \frac{N|\omega_k \cap c_j|}{|\omega_k||c_j|} \tag{185}$$

where $P(\omega_k)$, $P(c_j)$, and $P(\omega_k \cap c_j)$ are the probabilities of a document being in cluster ω_k, class c_j, and in the intersection of ω_k and c_j, respectively. Equation 185 is equivalent to Equation 184 for maximum likelihood estimates of the probabilities (i.e., the estimate of each probability is the corresponding relative frequency).

H is entropy as defined in Chapter 5 (page 5.3.2):

$$H(\Omega) \quad = \quad -\sum_k P(\omega_k) \log P(\omega_k) \tag{186}$$

$$= \quad -\sum_k \frac{|\omega_k|}{N} \log \frac{|\omega_k|}{N} \tag{187}$$

where, again, the second equation is based on maximum likelihood estimates of the probabilities.

$I(\Omega;\mathbb{C})$ in Equation 184 measures the amount of information by which our knowledge about the classes increases when we are told what the clusters are. The minimum of $I(\Omega;\mathbb{C})$ is 0 if the clustering is random with respect to class membership. In that case, knowing that a document is in a particular cluster does not give us any new information about what its class might be. Maximum mutual information is reached for a clustering Ω_{exact} that perfectly recreates the classes - but also if clusters in Ω_{exact} are further subdivided into smaller clusters (Exercise 16.7). In particular, a clustering with $K = N$ one-document clusters has maximum MI. So MI has the same problem as purity: it does not penalize large cardinalities and thus does not formalize our bias that, other things being equal, fewer clusters are better.

The normalization by the denominator $[H(\Omega) + H(\mathbb{C})]/2$ in Equation 183 fixes this problem since entropy tends to increase with the number of clusters. For example, $H(\Omega)$ reaches its maximum $\log N$ for $K = N$, which

ensures that NMI is low for $K = N$. Because NMI is normalized, we can use it to compare clusterings with different numbers of clusters. The particular form of the denominator is chosen because $[H(\Omega) + H(\mathbb{C})]/2$ is a tight upper bound on $I(\Omega;\mathbb{C})$ (Exercise 16.7). Thus, NMI is always a number between 0 and 1.

An alternative to this information-theoretic interpretation of clustering is to view it as a series of decisions, one for each of the $N(N-1)/2$ pairs of documents in the collection. We want to assign two documents to the same cluster if and only if they are similar. A true positive (TP) decision assigns two similar documents to the same cluster, a true negative (TN) decision assigns two dissimilar documents to different clusters. There are two types of errors we can commit. A (FP) decision assigns two dissimilar documents to the same cluster. A (FN) decision assigns two similar documents to different clusters. The

Rand index ( ) measures the percentage of decisions that are correct. That is, it is simply accuracy (Section 8.3 , page 8.3).

$$RI = \frac{TP+TN}{TP+FP+FN+TN}$$

As an example, we compute RI for Figure 16.4 . We first compute $TP+FP$. The three clusters contain 6, 6, and 5 points, respectively, so the total number of ``positives'' or pairs of documents that are in the same cluster is:

$$TP+FP = \binom{6}{2} + \binom{6}{2} + \binom{5}{2} = 40 \tag{188}$$

Of these, the x pairs in cluster 1, the o pairs in cluster 2, the ◇ pairs in cluster 3, and the x pair in cluster 3 are true positives:

$$TP = \binom{5}{2} + \binom{4}{2} + \binom{3}{2} + \binom{2}{2} = 20 \tag{189}$$

Thus, $FP = 40 - 20 = 20$.

FN and TN are computed similarly, resulting in the following contingency table:

	Same cluster	Different clusters
Same class	TP = 20	FN = 24
Different classes	FP = 20	TN = 72

RI is then $(20+72)/(20+20+24+72) \approx 0.68$.

The Rand index gives equal weight to false positives and false negatives. Separating similar documents is sometimes worse than putting pairs of dissimilar documents in the same cluster. We can use the *F measure* measuresperf to penalize false negatives more strongly than false positives by selecting a value $\beta > 1$, thus giving more weight to recall.

$$P = \frac{TP}{TP+FP} \qquad R = \frac{TP}{TP+FN} \qquad F_\beta = \frac{(\beta^2+1)PR}{\beta^2 P + R}$$

Based on the numbers in the contingency table, $P = 20/40 = 0.5$ and $R = 20/44 \approx 0.455$. This gives us $F_1 \approx 0.48$ for $\beta = 1$ and $F_5 \approx 0.456$ for $\beta = 5$. In information retrieval, evaluating clustering with F has

the advantage that the measure is already familiar to the research community.

Exercises.

- Replace every point d in Figure 16.4 with two identical copies of d in the same class. (i) Is it less difficult, equally difficult or more difficult to cluster this set of 34 points as opposed to the 17 points in Figure 16.4 ? (ii) Compute purity, NMI, RI, and F_5 for the clustering with 34 points. Which measures increase and which stay the same after doubling the number of points? (iii) Given your assessment in (i) and the results in (ii), which measures are best suited to compare the quality of the two clusterings?