Rand index

From Wikipedia, the free encyclopedia

The **Rand index**[1] or **Rand measure** (named after William M. Rand) in statistics, and in particular in data clustering, is a measure of the similarity between two data clusterings. A form of the Rand index may be defined that is adjusted for the chance grouping of elements, this is the **adjusted Rand index**. From a mathematical standpoint, Rand index is related to the accuracy, but is applicable even when class labels are not used.

Contents

Rand index

Definition

Given a set of n elements $S = \{o_1, \ldots, o_n\}$ and two partitions of S to compare, $X = \{X_1, \ldots, X_r\}$, a partition of *S* into *r* subsets, and $Y = \{Y_1, \ldots, Y_s\}$, a partition of *S* into *s* subsets, define the following:

- a, the number of pairs of elements in S that are in the same subset in X and in the same subset in Y
- b, the number of pairs of elements in S that are in different subsets in X and in different subsets in Y
- c, the number of pairs of elements in S that are in the same subset in X and in different subsets in Y
- d, the number of pairs of elements in S that are in different subsets in X and in the same subset in Y

The Rand index, R, is:[1][2]

$$R = \frac{a+b}{a+b+c+d} = \frac{a+b}{\binom{n}{2}}$$

Intuitively, $a + b$ can be considered as the number of agreements between X and Y and $c + d$ as the number of disagreements between X and Y.

Since the denominator is the total number of pairs, the Rand index represents the *frequency of occurrence* of agreements over the total pairs, or the probability that X and Y will agree on a randomly chosen pair.

Properties

The Rand index has a value between 0 and 1, with 0 indicating that the two data clusterings do not agree on any pair of points and 1 indicating that the data clusterings are exactly the same.

In mathematical terms, a, b, c, d are defined as follows:

- $a = |S^*|$, where $S^* = \{(o_i, o_j) | o_i, o_j \in X_k, o_i, o_j \in Y_l\}$
- $b = |S^*|$, where $S^* = \{(o_i, o_j) | o_i \in X_{k_1}, o_j \in X_{k_2}, o_i \in Y_{l_1}, o_j \in Y_{l_2}\}$
- $c = |S^*|$, where $S^* = \{(o_i, o_j) | o_i, o_j \in X_k, o_i \in Y_{l_1}, o_j \in Y_{l_2}\}$
- $d = |S^*|$, where $S^* = \{(o_i, o_j) | o_i \in X_{k_1}, o_j \in X_{k_2}, o_i, o_j \in Y_l\}$

for some $1 \leq i, j \leq n, i \neq j, 1 \leq k, k_1, k_2 \leq r, k_1 \neq k_2, 1 \leq l, l_1, l_2 \leq s, l_1 \neq l_2$

Adjusted Rand index

The adjusted Rand index is the corrected-for-chance version of the Rand index.[1][2][3] Though the Rand Index may only yield a value between 0 and +1, the adjusted Rand index can yield negative values if the index is less than the expected index.[4]

The contingency table

Given a set S of n elements, and two groupings or partitions (*e.g.* clusterings) of these points, namely $X = \{X_1, X_2, \ldots, X_r\}$ and $Y = \{Y_1, Y_2, \ldots, Y_s\}$, the overlap between X and Y can be summarized in a contingency table $[n_{ij}]$ where each entry n_{ij} denotes the number of objects in common between X_i and Y_j : $n_{ij} = |X_i \cap Y_j|$.

X\Y	Y_1	Y_2	$\ldots$	Y_s	**Sums**
X_1	n_{11}	n_{12}	$\ldots$	n_{1s}	a_1
X_2	n_{21}	n_{22}	$\ldots$	n_{2s}	a_2
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
X_r	n_{r1}	n_{r2}	$\ldots$	n_{rs}	a_r
Sums	b_1	b_2	$\ldots$	b_s	

Definition

The adjusted form of the Rand Index, the Adjusted Rand Index, is

$$\text{AdjustedIndex} = \frac{\text{Index} - \text{ExpectedIndex}}{\text{MaxIndex} - \text{ExpectedIndex}},$$ more specifically

$$ARI = \frac{\sum_{ij} \binom{n_{ij}}{2} - [\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}] / \binom{n}{2}}{\frac{1}{2} [\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}] - [\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}] / \binom{n}{2}}$$

where n_{ij}, a_i, b_j are values from the contingency table.

References

1. W. M. Rand (1971). "Objective criteria for the evaluation of clustering methods". *Journal of the American Statistical Association*. American Statistical Association. **66** (336): 846–850. doi:10.2307/2284239. JSTOR 2284239.
2. Lawrence Hubert and Phipps Arabie (1985). "Comparing partitions". *Journal of Classification* **2** (1): 193–218. doi:10.1007/BF01908075
3. Nguyen Xuan Vinh, Julien Epps and James Bailey (2009). PDF. "Information Theoretic Measures for Clustering Comparison: Is a Correction for Chance Necessary?" Check |URL= value (help) (PDF). *ICML '09: Proceedings of the 26th Annual International Conference on Machine Learning*. ACM. pp. 1073–1080. PDF (http://www.ima.umn.edu/~iwe n/REU/10.pdf).
4. http://i11www.iti.uni-karlsruhe.de/extra/publications/ww-cco-06.pdf

External links

- C++ implementation with MATLAB mex files (https://github.com/bjoern-andres/partition-comparison)

Categories: Summary statistics for contingency tables | Clustering criteria

---