WIKIPEDIA

Rand index

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.

Example clusterings for a dataset with the kMeans (left) and Mean shift (right) algorithms. The calculated Adjusted Rand index for these two clusterings is $ARI \approx 0.94$

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.

$\binom{n}{2}$ is calculated as $n(n-1)/2$.

Similarly, one can also view the Rand index as a measure of the percentage of correct decisions made by the algorithm. It can be computed using the following formula:

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

where TP is the number of true positives, TN is the number of true negatives, FP is the number of false positives, and FN is the number of false negatives.

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

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

Relationship with classification accuracy

The Rand index can also be viewed through the prism of binary classification accuracy over the pairs of elements in S. The two class labels are "o_i and o_j are in the same subset in X and Y" and "o_i and o_j are in different subsets in X and Y".

In that setting, a is the number of pairs correctly labeled as belonging to the same subset (true positives), and b is the number of pairs correctly labeled as belonging to different subsets (true negatives).

Adjusted Rand index

The adjusted Rand index is the corrected-for-chance version of the Rand index.[1][2][3] Such a correction for chance establishes a baseline by using the expected similarity of all pair-wise comparisons between clusterings specified by a random model. Traditionally, the Rand Index was corrected using the Permutation Model for clusterings (the number and size of clusters within a clustering are fixed, and all random clusterings are generated by shuffling the elements between the

fixed clusters). However, the premises of the permutation model are frequently violated; in many clustering scenarios, either the number of clusters or the size distribution of those clusters vary drastically. For example, consider that in K-means the number of clusters is fixed by the practitioner, but the sizes of those clusters are inferred from the data. Variations of the adjusted Rand Index account for different models of random clusterings.[4]

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.[5]

The contingency table

Given a set S of n elements, and two groupings or partitions (*e.g.* clusterings) of these elements, 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 : $n_{ij} = |X_i \cap Y_j|$.

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

Definition

The original Adjusted Rand Index using the Permutation Model is

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

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

See also

- Simple matching coefficient

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. arXiv:1704.01036 (https://arxiv.org/abs/1704.01036). doi:10.2307/2284239 (https://doi.org/10.2307%2F2284239). JSTOR 2284239 (https://www.jstor.org/stable/2284239).
2. Lawrence Hubert and Phipps Arabie (1985). "Comparing partitions". *Journal of Classification*. **2** (1): 193–218. doi:10.1007/BF01908075 (https://doi.org/10.1007%2FBF01908075).
3. Nguyen Xuan Vinh, Julien Epps and James Bailey (2009). "Information Theoretic Measures for Clustering Comparison: Is a Correction for Chance Necessary?" (http://www.jmlr.org/papers/volume11/vinh10a/vinh10a.pdf) (PDF). *ICML '09: Proceedings of the 26th Annual International*

Conference on Machine Learning. ACM. pp. 1073–1080.PDF (http://www.ima.umn.edu/~iwen/REU/10.pdf).
4. Alexander J Gates and Yong-Yeol Ahn (2017). "The Impact of Random Models on Clustering Similarity" (http://www.jmlr.org/papers/volume18/17-039/17-039.pdf) (PDF). *Journal of Machine Learning Research*. **18**: 1–28.PDF (http://www.jmlr.org/papers/volume18/17-039/17-039.pdf).
5. 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)