A FORMAL METHOD FOR SELECTING EVALUATION METRICS FOR IMAGE SEGMENTATION

Abdel Aziz Taha(1), Allan Hanbury(2), Oscar A. Jimenez del Toro(3)

MOTIVATION

✓ Dozens of evaluation metrics have been used to evaluate medical image segmentations in order to rank the systems that produced them.
✓ But each metric produces a ranking that may be different from the other rankings.

✓ Also, evaluation is dependent of the evaluation task.

✓ An effective identification of the state-of-the-art systems requires consistent selection of the evaluation metrics.
✓ There is a lack of a formal way to select evaluation metrics.

GOAL

A formal method for choosing the most suitable metric(s) from a metric pool to evaluate the quality of medical segmentations taking into account:

i. the segmentations being evaluated
ii. the segmentation task.

STRATEGY

Measuring the metric bias to the different properties of the segmentations being evaluated.

✓ This is achieved by analyzing how the average scores of subsets of the segmentations correlate when the subsets are selected (i) randomly and (ii) according to properties of the segmentations.

METHOD

I. Define a set of properties.
II. Measure the metric bias to each property.
III. Select the metric(s) with the lowest sum of biases over all properties.

(1) Random grouping
✓ Group the segmentations randomly into N subsets.
✓ Rank the subsets according to their average scores.
✓ Perform the same for each of the M metrics in a pool.

(2) Grouping by size
✓ Group the segmentations according to segment size into N subsets.
✓ Rank the groups according to their average scores.
✓ Perform the same for each of the M metrics in the pool.

(3) Ranking
✓ Now, you obtain M rankings from the random grouping (1) and M rankings from the grouping by size (2).
✓ Analyzing these rankings leads to the metric bias to the property used for grouping (segment size).

(4) Analyzing correlation
✓ The correlation between rankings of the random subsets reflects the nature of the metrics (base correlation).
✓ But the change in this correlation in the second case reflects the metrics bias to the property used (segment size) (biased correlation).

(5) Repeat, other properties
✓ Repeat (1) to (4) with other properties.
✓ Each time, the metric bias is the difference between the base correlation and the biased correlation.
✓ For each metric, sum the bias to all properties to find the overall bias.

(6) Select metric(s)
✓ The overall bias (bias sum) is an indicator of the metric suitability.
✓ Metrics with the least overall bias are the most suitable.
✓ Weights can be used to reflect the task specific issues, i.e. properties with more importance for the task are assigned accordingly higher weights than other properties.

RESULTS

We propose a formal method for ranking a set of metrics according to their suitability for evaluating a specific segmentation set, given a specific evaluation task.

The method has been tested against a manual ranking of 20 evaluation metrics, done by a medical expert.

As shown in the table, the ranking of metrics produced automatically by the proposed method correlates with the manual ranking by the medical expert. The correlation value is 0.607.

Complete experiment details can be found in [1].

ACKNOWLEDGMENT

✓ Thanks to Dr. Bijem H. Menze, ETH Zurich for providing the MRT brain segmentations from the MICCAI 12 BRATS challenge to be used as test data.
✓ The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement 318088 (VISCERAL).

CONTACT

(1) Abdel Aziz Taha
Vienna University of Technology, taha@ifs.tuwien.ac.at
(2) Allan Hanbury
Vienna University of Technology, hanbury@ifs.tuwien.ac.at
(3) Oscar A. Jimenez del Toro
Univ. of Applied Sciences Western Switzerland, oscar.jimenez@hevs.ch

REFERENCE