Data-Driven Rank Aggregation with Application to Grand Challenges

James Fishbaugh[1(✉)], Marcel Prastawa[2], Bo Wang[3], Patrick Reynolds[4], Stephen Aylward[4], and Guido Gerig[1]

[1] Tandon School of Engineering, New York University, New York, NY, USA
james.fishbaugh@nyu.edu
[2] Icahn School of Medicine, Mount Sinai, New York, NY, USA
[3] GE Global Research, Niskayuna, NY, USA
[4] Kitware Inc., Clifton Park, USA

Abstract. The increased number of challenges for comparative evaluation of biomedical image analysis procedures clearly reflects a need for unbiased assessment of the state-of-the-art methodological advances. Moreover, the ultimate translation of novel image analysis procedures to the clinic requires rigorous validation and evaluation of alternative schemes, a task that is best outsourced to the international research community. We commonly see an increase of the number of metrics to be used in parallel, reflecting alternative ways to measure similarity. Since different measures come with different scales and distributions, these are often normalized or converted into an individual rank ordering, leaving the problem of combining the set of multiple rankings into a final score. Proposed solutions are averaging or accumulation of rankings, raising the question if different metrics are to be treated the same or if all metrics would be needed to assess closeness to truth. We address this issue with a data-driven method for automatic estimation of weights for a set of metrics based on unsupervised rank aggregation. Our method requires no normalization procedures and makes no assumptions about metric distributions. We explore the sensitivity of metrics to small changes in input data with an iterative perturbation scheme, to prioritize the contribution of the most robust metrics in the overall ranking. We show on real anatomical data that our weighting scheme can dramatically change the ranking.

1 Introduction

In recent years, an increasing number of challenges are organized at international conferences in medical image analysis and computer vision as more and more imaging data sets are accessible [1–3]. These open and public challenges provide an ideal forum for researchers in both academia and industry to participate, with the goal of gaining a better understanding of the performance of various algorithms on a specific image analysis task. However, defining a set of metrics to evaluate a particular image analysis algorithm is a non trivial problem. In many cases, there are several metrics that need to be considered instead of

G. Gerig—Supported by grant NIH (NIBIB) 9R42MH106302-02.

M. Descoteaux et al. (Eds.): MICCAI 2017, Part II, LNCS 10434, pp. 754–762, 2017.
DOI: 10.1007/978-3-319-66185-8_85

only one. How to combine and balance the different metrics is an important issue, but usually the weights assigned to different metrics are chosen to be uniform or based on the organizers experience. This is not optimal, as changing to a different set of weights can lead to a different rank of all the participants. From the perspective of a participant, a given rank ordering declares a winner and awards prize and prestige. However, the spirit of a challenge is to better understanding the benefits and drawbacks of the various algorithms rather than produce a leaderboard ordering. Challenge results may have further reaching effects, for example on commercial product development, method refinement, or inspire the design of new algorithms.

In [4], the STAPLE algorithm takes a collection of segmentations as input and estimates true segmentations as well as a measure of the performance level of each input, which enables assessment of the performance of automatic image segmentation methods. In [5], an evaluation of 14 nonlinear deformation algorithms was conducted by using three independent statistical analyses with overlap, volume similarity, and distance measures. A set of measures for the validation of diffusion tensor imaging (DTI) tractography is proposed in [6], and applied the proposed methods to evaluate atlas building-based tractography. The BRATS challenge was analyzed in [1] to explore the reason some algorithms worked better than others. The work of [7] proposes ordering metrics by their bias towards a set of defined segmentation properties. While [7] helps inform which metrics to include, it does not give a solution for combining metrics.

In this paper we combine a data-driven, unsupervised rank aggregation scheme with a perturbation based analysis of metric sensitivity to automatically compute weights for a set of metrics. Our method does not require normalization of metrics and makes no assumption about the distribution of metric values. Rather, the estimation of weights and corresponding rank ordering is determined entirely by the data and the *specific* image analysis task. We show on real anatomical data that by applying the proposed scheme, the final rank order may dramatically change, a result we hope will raise awareness in the community about shortcomings of current ad-hoc evaluation methods. Results demonstrate that the iterative procedure results in weights that reflect contributions of each metric in a plausible way, thus providing improved insight into overall rank aggregation. Our methodology provides transparency on aggregation that may help future challenge organizers to evaluate the best set of metrics beforehand based on existing data. We also advocate for our method as an exploratory tool, as the resulting weight for each metric, each usually representing a different aspect of similarity (overlap, surface distance, sensitivity to outliers, etc.), provides essential information for algorithm assessment. Finally, we reiterate that the motivation of our work is to better understand the performance of various algorithms, not to produce a de facto leaderboard ordering.

2 Methodology

The principal behind this rank aggregation scheme, introduced in [8], is that reliable metrics rank submissions in a similar manner. Metrics that produces an

ordering that tend to agree with a collection of other metrics will be given a higher weight. Conversly, metrics that produce inconsistent rankings will receive lower weights. Central to the concept of consistent orderings is the stability of metrics. We propose to measure sensitivity of metrics to small perturbations of the input data. The insight is that metrics that are robust to small changes to input should receive a higher weight, and those metrics that produce different orderings under perturbations should receive lower weights.

2.1 Rank Aggregation

Let $G = (g_1, ..., g_{N_g})$ represent a number of ground truth data for N_g tasks. For example, these could be segmentations for $N_g = 5$ different cases. A single ground truth could be assembled from several expert sources using label fusion, STAPLE [4], or a custom algorithm [1]. Let $\mathbf{X} = (X_1, ..., X_{N_S})$ be N_S submissions to the challenge, where each submission $X_i = (x_1^i, ..., x_{N_g}^i)$ represents a set of N_G items to be directly compared to the set of ground truth data G. Let $\mathbf{M} = (M_1, ..., M_{N_M})$ represent a collection of metrics, where for a given $g \in G$ and $x \in X_i$, a metric $M(g, x)$ returns a scalar value. Let $\mathbf{R} = (R_1, ..., R_{N_M})$ be the ordinal rankings (ranking functions) corresponding to metrics $\mathbf{M}$ evaluated on all submissions. For example, R_i is the ordinal ordering (1, 2, ..., N_S) of all the submissions under metric function M_i.

We require an aggregate ranking function $A(\mathbf{R}, G, \mathbf{X}) = \sum_{i=1}^{N_M} w_i R_i(G, \mathbf{X})$ as linear combination of orderings R_i (given by metric M_i) to produce an overall ranking. This linear combination is parameterized by weights $W = (w_1, ..., w_{N_M})$ which can be thought of as a probability density function ($\sum_i^{N_M} w_i = 1$).

For a given submission item x and its corresponding ground truth g, we can compute the average ranking across all metrics defined as

$$\mu(g, x) = \frac{\sum_{i=1}^{N_M} R_i(g, x)}{N_M}. \tag{1}$$

The mean value $\mu(g, x)$ value can then be used to capture the variance of any individual ranking R_i under metric M_i by $\sigma_i(g, x) = [R_i(g, x) - \mu(g, x)]^2$. A small value of σ_i suggests that metric M_i produces an ordering in agreement with the other metrics and should be given a higher weight, while large values of σ_i represent disagreement from the majority and will be receive a lower weight. Note that the computation of mean and variance here is with respect to the *orderings* given by the metric, not the metrics themselves. That way, no normalization of metrics is required, and no assumptions about the distribution of metric values is assumed.

We can then pose this as an optimization problem, to find weights which minimize σ_i over all submissions:

$$\underset{W}{\operatorname{argmin}} \sum_{g \in G} \sum_{x \in \mathbf{X}} \sum_{i=1}^{N_M} w_i \sigma_i(g, x) \tag{2}$$

Algorithm 1. Unsupervised Rank Aggregation

Input: $\mathbf{R}[G, \mathbf{X}, \mathbf{M}]$ ▷ Values for all ground truth, submissions, and metrics
$\mathbf{w} \leftarrow 1/N_M$ ▷ Initialize uniform weights
while ($i <$ maxIterations) **and** (**not** converged) **do**
  for all g **in** G **do**
    for all x **in X do**
      $\mu(g,x) \leftarrow \frac{\sum_{i=1}^{N_M} R_i[g,x,M_i]}{N_M}$ ▷ Compute average ranking for this item
      for all M **in R do**
        $\nabla_{w_M} \leftarrow (\mathbf{R}[g,x,M] - \mu(g,x))^2$ ▷ Compute gradient
        $w_M^i \leftarrow \frac{w_M^{i-1}\exp(-\lambda\nabla_{w_M})}{\sum_{j=1}^{N_M} w_j^{i-1}\exp(-\lambda\nabla_{w_M})}$ ▷ Update weight
  $i \leftarrow i+1$
return w

with the constraint that $\sum_i^{N_M} w_i = 1$ and $0 \leq w_i \leq 1$. The gradient with respect to a given w_i is

$$\nabla_{w_i} = [R_i(g,x) - \mu(g,x)]^2, \tag{3}$$

which can be used to derive a gradient descent scheme [8], summarized in Algorithm 1.

Example: Consider a synthetic example where the true ranking is {1, 2, ..., 15}. We have 10 total ranking functions of varying accuracy, summarized in Table 1. The first three ranking functions give the correct ordering. The next three ranking functions return a list close to the correct ordering, created with 5 random swaps of adjacent items. The next two ranking functions are unreliable, with 5 random swaps of any two items in the ordering. The final two ranking functions are purely random orderings.

The method detects that the first three ranking functions are the most consistent and assigns the highest weight of 0.176. The next three ranking functions are less consistent and receive a slightly lower weight of 0.168, 0.146, and 0.146. The next two ranking functions are inconsistent and are given weights 0.002 and 0.008. The final two random ranking functions are correctly given weights of 0. With the estimated weights, the correct ordering of {1, 2, ..., 15} is produced. The ranking with uniform weights is the incorrect ordering {3, 2, 1, 4, 5, 7, 8, 6, 10, 9, 12, 11, 13, 15, 14}.

Correlation: Inherent to the task of evaluating segmentation is the problem of metric selection and correlation. Indeed, similarity metrics are often highly correlated, or nearly identical in the case of dice and Cohen's kappa. As our method favors metrics which are in agreement with other metrics, we must address metric correlation. Guided by the work of [9], we carefully choose a collection of metrics to capture a wide range of metric properties while limiting the use of highly or perfectly correlated metrics. We include overlap based metrics dice, global consistency error, sensitivity, and specificity; surface based metric Hausdorff distance (95th percentile); information theoretic measure mutual information; and volume based measure volumetric similarity.

Table 1. Synthetic example where true ranking is {1, 2, ..., 15}. Ranking functions 1,2,3 are perfect, 4,5,6 contain small errors, 7,8 contain large errors, and 9 and 10 are random.

Ranking functions	Ordering	Weights
1	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15	0.176
2	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15	0.176
3	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15	0.176
4	2, 1, 3, 5, 6, 4, 8, 7, 10, 9, 11, 12, 13, 14, 15	0.168
5	2, 1, 4, 5, 3, 6, 7, 8, 9, 10, 11, 12, 14, 15, 13	0.146
6	1, 3, 2, 4, 5, 6, 7, 8, 10, 11, 9, 12, 13, 14, 15	0.146
7	13, 2, 3, 4, 5, 6, 8, 1, 9, 10, 12, 7, 11, 14, 15	0.002
8	13, 2, 3, 4, 5, 6, 8, 1, 9, 10, 12, 7, 11, 14, 15	0.008
9	11, 8, 2, 6, 1, 15, 9, 4, 12, 3, 7, 5, 14, 13, 9	0
10	6, 15, 1, 8, 2, 12, 14, 5, 4, 3, 10, 9, 7, 13, 11	0

2.2 Assessing Stability with Perturbations

The traditional domains for rank aggregation, such as elections or meta-search, deal with ordinal rankings from the onset. That is to say, there is a collection of rankings provided by different ranking functions, but the inner workings of the ranking functions are either not available or not defined. In the meta-search example, the ranking functions are often proprietary, and for elections the ranking functions are based on personal preference. In these situations, the only recourse is to deal with the rankings directly.

In this work, we have the unique opportunity to systematically explore the ranking functions themselves. We propose to do this by assessing the stability and robustness of metrics by small perturbations to the input data. The intuition is that metrics that are robust to small perturbations provide more consistent rank ordering and should receive higher weights. Conversely, a metric where a small change in input data leads to a large change in the resulting ordering should be considered too sensitive to reliably discriminate differences, and should receive a lower weight. Combining perturbations with rank aggregation allows metric weights to reflect the sensitivity of the metrics on the *specific* image analysis task, completely determined by the data.

The method works by iteratively applying perturbations to ground truth data, and re-estimating weights using the scheme in Sect. 2.1, while keeping a running average of estimated weights. The necessary component is a method (or methods) to perturb ground truth data. A perturbation method could be general purpose deformations such as rigid transformations, or could be a custom algorithm designed with expert knowledge to accurate mimic anatomical variability. Whatever the method used to make small modifications to ground truth, the key is to produce a number of unique perturbations to fully probe each metric for reliability.

Algorithm 2. Rank Aggregation with Perturbations

$\mathbf{w} \leftarrow 0$ ▷ Initialize weights
for $(i = 1; i <$ numPerturbations$; i{+}{+})$ **do**
  for all $g \in G$ **do**
    $\hat{g} \leftarrow$ `PerturbData`(g) ▷ Choose a perturbation method
    for all $X \in \mathbf{X}$ **do**
      $x \leftarrow x \in X$ corresponding to g
      for all $M \in \mathbf{M}$ **do**
        $R[\hat{g}, x, M] \leftarrow M(\hat{g}, x)$ ▷ Ranking based on metric M
  $\mathbf{w} \leftarrow \mathbf{w} +$ `RankAggregation`$(\mathbf{R})$ ▷ Compute weights (alg. 1)
$\mathbf{w} \leftarrow \mathbf{w}/$numPerturbations ▷ Final weights are averaged over all perturbations

3 Experimental Validation

Data: We test our rank aggregation scheme on an artificial challenge to segment the corpus callosum, a flat bundle of fibers which connect the left and right hemisphere of the brain. The 2D contour of the corpus callosum is clearly visible in mid-sagittal slices from 3D brain MRI. Our data consists of 10 unique subjects (2D sagittal slices) that are repeated 3 times each to form a dataset of 30 images, where the image ordering is randomly permuted. Submitters were asked to manually outline the 30 corpus callosum structures using `itksnap` [10], without knowledge that it was 10 subjects repeated 3 times each. In total, 6 submitters provided outlines, which can be considered 18 unique submissions by taking into account the repeated nature of the data. For evaluation, ground truth segmentations were obtained by a deformable active contour model [11]. An example corpus callosum segmentation is shown in Fig. 1.

Metrics: To evaluate each submission with respect to ground truth, we employ several metrics discussed in Sect. 2. We include dice, global consistency error, sensitivity, and specificity; Hausdorff distance (95th percentile); mutual information, and volumetric similarity. The metrics were chosen to capture a wide range of metric properties while limiting the use of highly or perfectly correlated metrics, as shown in [9].

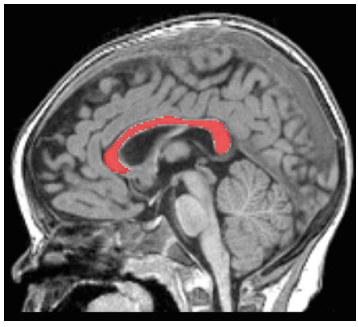

Fig. 1. Corpus callosum shown in red.

Perturbations: We implement 3 perturbation methods. For modeling linear transformations, we use rigid perturbations with a specified amount of random translation, rotation, and scaling. To model submissions who might over or under segment, we use morphological perturbations which randomly iterate between dilation and erosion. Finally, to model nonlinear differences from ground truth, we use B-spline perturbations with randomness

controlled through a Gaussian distributed random sampling for B-spline parameters. For each iteration, a random perturbation method is chosen.

Results: We explore our rank aggregation with perturbations framework by considering 18 submissions of the 10 corpus callosum segmentation tasks. The submissions are named A–F with suffix denoting the 3 repeated segmentation tasks. The left side of Table 2 shows the final overall ranking using naive uniform weights for each metric. Our proposed method estimates weights: dice = 0.33, mutual information = 0.21, specificity = 0.20, volumetric similarity = 0.12, Hausdorff distance = 0.10, sensitivity = 0.03, and global consistency error = 0.01. The overall ranking under the computed weights is shown on the right of Table 2. It is interesting to note that the estimated weights dramatically changed the overall order as compared to uniform weights. In this case, global consistency error and sensitivity produce inconsistent orderings under perturbations and receive a low weight. The distribution of weights has the potential to provide important insight into why certain algorithms perform well on a given medical imaging task, which is the true spirit of grand challenges. Such feedback may serve to inform algorithm refinement, or help steer new algorithm development. For example, we may gain insight that a particular problem is better solved by a method based on intensities, contrast, shape models, or physical models.

We also explore how the number of perturbations influences the final estimated weights, as well as the magnitude of perturbations. For "small" perturbations, we set rigid scale parameters to 2 pixels translation, 5° rotation, and

Table 2. For the corpus callosum challenge, the overall ranking using uniform weights compared to the weights estimated from rank aggregation with perturbations.

Rank	Uniform Weights	Estimated Weights
1	Submission C2	Submission C2
2	Submission C1	Submission F3
3	Submission C3	Submission C3
4	Submission F3	Submission C1
5	Submission A1	Submission D3
6	Submission E3	Submission E3
7	Submission B2	Submission F1
8	Submission E2	Submission A1
9	Submission B1	Submission B1
10	Submission F1	Submission E2
11	Submission E1	Submission B2
12	Submission A2	Submission A3
13	Submission A3	Submission F2
14	Submission D3	Submission A2
15	Submission B3	Submission D2
16	Submission D1	Submission D1
17	Submission F2	Submission E1
18	Submission D2	Submission B3

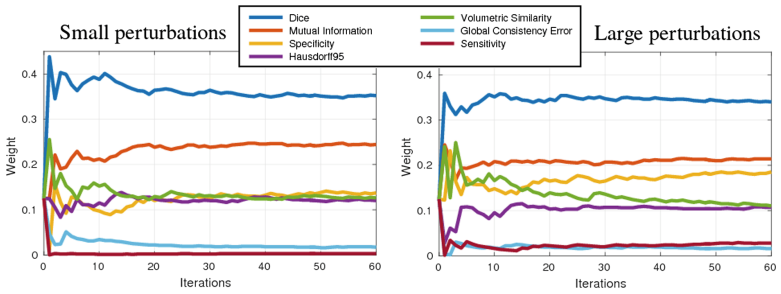

Fig. 2. Evolution of weights vs. the number of iterations of perturbations.

5% scale, morphology parameters to 5 iterations, and B-spline variance to 2.0. For "large" perturbations, we set rigid scale parameters up to 10 pixels translation, 30° rotation, and 50% scale, morphology parameters to 10 iterations, and B-spline variance to 15.0. Figure 2 summarizes the results of these experiments. For this experiment, large perturbations seem to provide more separation between metric weights, particularly increasing the weight of specificity and increasing the relative importance of dice. We also observe faster convergence to stable weights under small perturbations, as large perturbations introduce more variability in orderings.

4 Conclusion

We have presented a method to automatically calculate weights for a set of metrics which probes the sensitivity of the metrics by exploring changes in rank due to perturbations to input data. Our method is completely data-driven, requiring no metric normalization procedures. We showed how our estimated weights can result in a vastly different ordering compared to uniform weighting. This has the potential to better inform organizers about the results, and provide additional insight into the performance of competing algorithms. For example, the distribution of weights and corresponding ranking changes may provide a clue that a particular problem is better solved by a method based on intensities, contrast, shape models, or physical models. Correlation is currently handled by careful selection of metrics. What remains is to automatically select the best metrics in addition to their weight, perhaps by integrating the work of [7]. Future work will explore and validate our method on data from a public challenge.

References

1. Menze, B.H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., Burren, Y., Porz, N., Slotboom, J., Wiest, R., et al.: The multimodal brain tumor image segmentation benchmark (BRATS). IEEE Trans. Med. Imaging **34**(10), 1993–2024 (2015)

2. Pujol, S., Wells, W., Pierpaoli, C., Brun, C., Gee, J., Cheng, G., Vemuri, B., Commowick, O., Prima, S., et al.: The DTI challenge: toward standardized evaluation of diffusion tensor imaging tractography for neurosurgery. J. Neuroimaging **25**(6), 875–882 (2015)
3. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L.: ImageNet large scale visual recognition challenge. Int. J. Comput. Vis. **115**(3), 211–252 (2015)
4. Warfield, S., Zou, K., Wells, W.: Simultaneous truth and performance level estimation (staple): an algorithm for the validation of image segmentation. IEEE Trans. Med. Imaging **23**(7), 903–921 (2004)
5. Klein, A., Andersson, J., Ardekani, B.A., Ashburner, J., Avants, B., Chiang, M.C., Christensen, G.E., Collins, D.L., Gee, J., Hellier, P., et al.: Evaluation of 14 nonlinear deformation algorithms applied to human brain MRI registration. Neuroimage **46**(3), 786–802 (2009)
6. Gouttard, S., Goodlett, C.B., Kubicki, M., Gerig, G.: Measures for validation of DTI tractography. In: SPIE Medical Imaging, ISOP, p. 83140J (2012)
7. Taha, A.A., Hanbury, A., del Toro, O.A.J.: A formal method for selecting evaluation metrics for image segmentation. In: ICIP, pp. 932–936. IEEE (2014)
8. Klementiev, A., Roth, D., Small, K.: An unsupervised learning algorithm for rank aggregation. In: Kok, J.N., Koronacki, J., Mantaras, R.L., Matwin, S., Mladenič, D., Skowron, A. (eds.) ECML 2007. LNCS, vol. 4701, pp. 616–623. Springer, Heidelberg (2007). doi:10.1007/978-3-540-74958-5_60
9. Taha, A.A., Hanbury, A.: Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool. BMC Med. Imaging **15**(1), 29 (2015)
10. Yushkevich, P., Piven, J., Cody, H., Ho, S., Gee, J.C., Gerig, G.: User-guided level set segmentation of anatomical structures with ITK-SNAP. NeuroImage **31**, 1116–1128 (2005)
11. Vachet, C., Yvernault, B., Bhatt, K., Smith, R.G., Gerig, G., Hazlett, H.C., Styner, M.: Automatic corpus callosum segmentation using a deformable active Fourier contour model. In: SPIE Medical Imaging, ISOP, p. 831707 (2012)