A FORMAL METHOD FOR SELECTING EVALUATION METRICS FOR IMAGE SEGMENTATION.

Abdel Aziz Taha, Allan Hanbury

Vienna University of Technology

Oscar A. Jimenez del Toro

Univ. of Applied Sciences Western Swizerland

ABSTRACT

Evaluating the quality of segmentations is an important process in image processing, especially in the medical domain. Many evaluation metrics have been used in evaluating segmentation. There exists no formal way to choose the most suitable metric(s) for a particular segmentation task and/or particular data. In this paper we propose a formal method for choosing the most suitable metrics for evaluating the quality of segmentations with respect to ground truth segmentations. The proposed method depends on measuring the bias of metrics towards/against the properties of the the segmentations being evaluated. We firstly demonstrate how metrics can have bias towards/against particular properties and then we propose a general method for ranking metrics according to their overall bias. We finally demonstrate for 3D medical image segmentations that ranking produced using metrics with low overall bias strongly correlate with manual rankings done by an expert.

Index Terms— image segmentation; evaluation metrics; selection

1 Introduction

1.1 The need to understand metrics: Many evaluation metrics for image segmentation have been introduced; most researchers choose the evaluation metrics arbitrarily or according to their popularity. Investigating metrics would help researchers to better understand them and help companies and stakeholders to save effort and time reaching optimal systems [1]. A poorly defined metric may lead to inaccurate conclusions like selecting suboptimal models when comparing the performance of classifiers [2].

Many researchers have investigated the drawbacks of particular metrics given particular properties of the data being classified. As a special case of classification, image segmentation is also affected by these drawbacks. The following are some examples: Hausdorff distance is very sensitive to noise and least squares based evaluation methods are very sensitive to outliers [3]. Mutual information doesn't utilize spatial information inherited in images because only voxel relationships are considered but not the neighborhoods [4]. Information theoretical measures have a non-convergent baseline which depends on the ratio between the number of data points and the number of classes. Therefore this class of measure needs chance correction [5]. Commonly used measures (precision, recall and F-measures) are biased and don't consider the level of chance [6]. Choosing evaluation metrics is very important and application-dependent; when evaluating imbalanced datasets, the metric choice is not obvious [2]. Metrics have different properties with respect to their correlation with user satisfaction criteria and their ease of interpretation [7]. Benhabiles et al. [8] validated 250 automatic segmentations against their corresponding ground truth segmentations using four different evaluation metrics. The results were then compared with manual ratings from 40 human observers. They found that the correlations between the ranking based on the manual ratings and the rankings based on the evaluation metrics vary between 30% and 80% depending on the used metric.

Research in the last decades generally results in the relative system improvement achieved becoming smaller and smaller. As a result, sensitivity and fidelity of evaluation metrics become increasingly critical. When improvements are small, metrics with high sensitivity are needed to measure small but real improvements and also with high fidelity to distinguish between improvements based on user preferences and improvements resulting from biased relevance judgments [9] [10].

1.2 Problem definition and notations: In this paper, we propose a formal method for selecting the most suitable metrics to evaluate image segmentation depending on the data being segmented and the goal of the segmentation task. The method is primarily based on two facts: the first is that effectiveness metrics can be biased towards or against properties of the images being segmented, meaning that particular metrics over-

penalize or over-reward segmentations given particular properties [4] [6] [11] [2] [3]. The second fact is that selecting the best evaluation metrics can be subject to the segmentation goal which means that the bias towards/against a particular property of the data can be differently important depending on the segmentation goal [8] [7]. To meet the context dependency, the proposed method allows individual weighting of the influence of each property according to its importance in case this is known, which increases the effectiveness of the method.

The problem to be solved in this paper can be formulated as follows: given a set of metrics $M = \{M_1, M2, ..., M_r\}$, a set of image segmentations $C = \{C_1, C_2, ..., C_k\}$, then the task is to rank the metrics in M according to their suitability for evaluating the quality of the segmentations in C provided that for each segmentation there exists a ground truth segmentation.

The proposed method is general and can be applied to select evaluation metrics for all types of segmentations. However, for simplicity, we will consider only the crisp segmentation task in this paper to present and formulate the method. In particular, we will be analyzing and testing the method using a special type of segmentation, namely medical volume segmentation e.g. magnetic resonance images (MRI) where voxels (3D pixels) are either assigned or not to a given class (segment) e.g. an organ or a tumor.

2 Related Work

Jin et al. [12] established a formal method for comparing two different measures and introduced two criteria for formal comparison of the goodness of evaluation metrics, namely the degree of consistency (DoC) and degree of discriminancy (DoD). Applying these criteria, they showed theoretically and empirically that AUC is a better measure than accuracy in evaluating the performance of classifiers. [13] [14] applied formal constraints based on axiometry to compare and judge evaluation metrics depending on the grade of satisfaction of these constraints. Busin et al. [15] used axiometrics to define a formal and general notation that fits any effectiveness metric. Based on this notation, they proposed several axioms that should be satisfied by an effectiveness metric. They used these axioms as criteria to evaluate metrics. All these papers deal with the problem only from a theoretical axiometrical point of view without taking into account the classification goal and the nature and properties of data being classified.

Sakai [11] proposed a method for evaluating evaluation metrics by measuring their sensitivity using Bootstrap Hypothesis Tests, and used this method in comparing seven evaluation metrics. They negate the belief that commonly used evaluation measures are equally reliable. Fatourechi et al. [2] proposed a framework based on Desired Region of Operation (DROP) for selecting the best evaluation metric for evaluating imbalanced classifications. Sakai [16] provided comparisons between metrics depending on the sensitivity and stability using the Voorhees/Buckley swap method [17]. All these papers lack generality because they are methods designed either for specific metrics or for specific metric properties.

3 Proposed Methods

We propose a method for choosing the most suitable metric for evaluating image segmentation. In Sections 3.1 to 3.3, the method is described and discussed formally. Then this formal description is explained in a step by step demonstration with a real example in Section 4, which also provides an experimental evaluation of the method.

Given a set of effectiveness metrics M and a set of segmentations C, each of the segmentations is evaluated against its ground truth segmentation using all metrics to obtain a ranking per metric. Now, choosing the most suitable metric goes in two main steps: (i) Constructing different partitions on the segmentation set C and ranking the subsets of each partition according to their average quality regarding each metric. (ii) Inferring the metric bias from the rank correlations across all partitions and all metrics. In the following, each of the steps is described in more details.

3.1 Constructing partitions and rank structure:

1. For each metric $m \in M$, evaluate each of the segmentations $x \in C$ against its ground truth segmentation to get the score matrix s where $s(x, m)$ is the score of segmentation x measured by metric m.

2. For each metric $m \in M$, assign each segmentation $x \in C$ a rank depending on its score to get the rank matrix r where $r(x, m)$ is the rank of segmentation x measured by metric m.

3. Define a set F of t segmentation properties. These can be any features thought to impact metrics e.g. class imbalance, number of segments, segment size, noise, deviation, shape signatures, sphereness, boundary smoothness, resolution, moments, etc. Furthermore, features can also be score-dependent e.g. precision and recall for utilizing trade-off i.e. for evaluation that tends to reward precision on cost of recall and vice versa. If no features are known to impact metrics, simply use all available features.

4. Construct t different partitions on the segmentation set C, each partition according to one feature from F, i.e. according to the grade of occurrence of the feature in the segmentations. One gets the set of partitions $P = \{P_1, .., P_t\}$. Each partition should have the same number of subsets s. The function $P_{ij}(x)$ assigns the segmentation x to the subset j according to partition i.

$$P_{ij}(x) = \begin{cases} 1 & x \in \text{ subset } j \text{ of partition } i \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

5. Construct t random partitions $\check{P} = \{\check{P}_1, .., \check{P}_t\}$ by randomly assigning segmentations to s equal subsets in each partition. The function $\check{P}_{ij}(x)$ that assigns a segmentation x to the subset j of the random partition i is defined by

$$\check{P}_{ij}(x) = \begin{cases} 1 & x \in \text{ subset } j \text{ of random partition } i \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

6. For each metric $m \in M$, for each partition $i \in P$, rank the subsets j according to the average of the individual ranks in each subset using the rank function

$$s_{mi}(j) = (\sum_{P_{ij}(x)=1} r(m, x))/n_{ij} \quad (3)$$

where $x \in C$ are the individual segmentations and n_{ij} is the number of segmentations in the subset j of Partition i. Now, use the rank averages from Equation 3 to compute the rank structure $R = R(i, j, m)$ that gives the rank of subset j of partition i measured by metric m according to descending rank average. Analogously, $\check{R} = \check{R}(i, j, m)$ gives the rank of subset j of the random partition i measured by metric m. Note that ranking the subsets using the averages of the individual ranks in each subset is a ranking method inspired by the Mann-Whitney-Wilcoxon (MWW) test [18]. This is because straightforwardly computing the ranks from score averages is sensitive to outliers and may produce unreasonable rankings if the scores are not normally distributed [19].

3.2 Inferring metric bias: Now, the rank structures $\check{R}$ (rankings of the random partitions) and R (rankings of the non-random partitions) provide a statistical basis to infer metric bias by analyzing how rankings of the different metrics and different partitions are correlated. The analysis is primarily based on comparing two correlations: the average of the rank correlations given the random partitions $\check{R}$ (we will call this correlation the base correlation $\check{K}$) and the rank correlation given a particular partition R (we will call this correlation the biased correlation K). They are given by

$$\check{K}(m_t) = \frac{1}{|\check{P}|.|M|} \sum_{i \in \check{P}} \sum_{m \in M} r[\check{R}(i, ., m_t), \check{R}(i, ., m)] \quad (4)$$

$$K(m_t, p) = \frac{1}{|M|} \sum_{m \in M} r[R(p, ., m_t), R(p, ., m)] \quad (5)$$

where m_t is the metric being evaluated, p is a given partition, and $r(x1, x2)$ is the Pearson's correlation coefficient between the rankings $x1$ and $x2$ (the point denotes all possible values, e.g. $R(p, ., m)$ means all possible subset ranks in partition p measured by metric m)

Now, we define the overall bias of metric m_t to be the average of the absolute correlation change $B(m_t)$ which is given by

$$B(m_t) = \frac{1}{|P|} \sum_{i \in P} abs[K(m_t, i) - \check{K}(m_t)] \quad (6)$$

Finally, the metrics in M are ranked according to their overall bias, where the metric with the lowest bias is the most suitable.

3.3 Discussion: To understand the key idea, let's think about the following two cases: Case 1, partitioning the segmentations randomly. Case 2, partitioning the segmentations according to a particular property (e.g. class imbalance in the underlying segmentations). Given a particular metric m, the base correlation $\check{K}(m)$ given by Equation 4 (related to Case 1) depends on the nature of the metric and is not affected by the properties of the segmentations, since the partition is random. Now, if this correlation changes when we consider Case 2 (i.e. biased correlation K given by Equation 5), then the change is caused by the impact of the property used for partitioning (in this case class imbalance) on the metric and therefore it characterizes the bias of the metric towards/against this property. If many partitions (many properties) are used, then the sum of the correlation differences is a measure of the overall bias for the given metric.

4 Experiments

In this section, the proposed method is demonstrated and tested with a real example, namely a set of 229 automatic brain tumor segmentations (MRI 3D volumes) from the BRATS2012 challenge[1]. The segmentations correspond to 47 medical cases and were produced by

[1]MICCAI 2012 Challenge on Multimodal Brain Tumor Segmentation, www2.imm.dtu.dk/projects/BRATS2012

five different algorithms participating in BRATS challenge. To build the rank structure (described in Section 3.1), all segmentations were evaluated against their ground truth segmentations using 18 metrics (listed in Table 1) to get the score matrix s (Step 1). Then global ranks were calculated from scores to get a ranking per metric r (Step 2). A set of 7 properties, namely segment size, noise, class imbalance, connected component count, point variance, sphereness, and recall, was defined (Step 3). Now, 7 partitions of the segmentations were constructed each time using one of the defined properties. Each consists of 10 subsets with $\frac{229}{10} \approx 22$ segmentations (Step 4). A random partition with 10 equal subsets was constructed (Step 5). For each partition, the subsets were ranked using the sum of individual global ranks r to get 18 rankings per partition (126 rankings in total). The random partition was also ranked to get 18 rankings (Step 6). To infer metric bias, Equations 4, 5, and 6 in Section 3.2 were applied to the resulting rank structure. The result of this step is a metric list ranked according to bias.

To validate the suitability ranking produced by the proposed approach, a manual ranking done by a radiology expert was used: for each medical case, the five corresponding segmentations were ranked by their quality from a medical point of view (we call these the manual rankings). Analogously, for each medical case, 18 rankings of the five segmentations were produced each time using one of metrics (we call these the metric rankings). The average correlation between manual rankings and metric rankings was computed for each metric and finally the metrics were sorted according to this average correlation. The resulting metric ranking (Table 1 column 'manual') was used as a ground truth suitability ranking of the metrics to validate the automatic ranking. Table 1 column 'automatic' contains for each metric the bias (Equation 6) and the corresponding suitability rank computed according to ascending bias. A moderate to strong correlation between the two rankings can be observed. The six best metrics are the same in both rankings. This correlation shows that metrics with low bias produce rankings that are more correlated to manual rankings than others.

metric	manual		automatic	
	correl.	rank	bias	rank
Cohen's Kappa	0.818	1	33.5	2
Adjusted Rand Index	0.818	1	33.1	1
Interclass Correlation	0.818	1	33.5	2
Probabilistic distance	0.802	2	34.7	5
Dice	0.800	3	33.6	3
Average Distance	0.798	4	33.9	4
Accuracy	0.791	5	64.0	14
Rand Index	0.791	5	64.0	14
Variation of Inform.	0.791	6	62.0	13
Mutual Information	0.753	7	46.5	12
Mahalanobis Distance	0.701	8	37.7	7
Global Consistency Err.	0.670	9	69.8	15
Hausdorff Distance	0.663	10	35.5	6
Area u. curve (AUC)	0.647	11	42.0	8
Sensitivity	0.615	12	44.4	10
Precision	0.608	13	44.5	11
Volumetric Similarity	0.590	14	43.6	9
Specificity	0.398	15	78.6	16
Correl. btw. manual & automatic ranking				0.607

Table 1. Manual and automatic metric suitability rankings. In column 'manual', the average correlation between metric rankings and the manual rankings as well as corresponding suitability ranks according to decending correlation. In column 'automatic', the metric bias calculated automatically by the proposed method as well as the ranks according to ascending bias (detailed data and results available in [20])

5 Conclusion and future work

For evaluating segmentations, metrics can be chosen according to their bias (Equation 6) toward/against the properties of the segmentations being evaluated. Test results show that the ranking produced by metrics with low bias generally have higher correlation with manual ranking than rankings produced by other metrics. In future work, the method will be tested with segmentations of other types and validated against rankings from different experts. A further issue to be investigated in future work is the influence of weighting the properties in Equation 6 on the metric suitability ranking, if it is known that particular properties are more/less important for the segmentation goal. For example, the manual ranking used to validate this method is done by a radiologist who may emphasise recall on cost of precision to assure that the tumor is completely removed. In this case weighting the recall and precision properly could improve the result.

6 Acknowledgments

Thanks to Dr. Bjoern H. Menze, ETH Zurich for providing the MRI brain segmentations from 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 318068 (VISCERAL).

7 References

[1] T.H.J.M. Peeters, P.R. Rodrigues, A. Vilanova, and B.M ter Haar Romeny, "Analysis of distance/similarity measures for diffusion tensor imaging," in *Visualization and Processing of Tensor Fields: Advances and Perspectives*. Springer, Berlin, 2008.

[2] Mehrdad Fatourechi, Rabab K. Ward, Steven G. Mason, Jane Huggins, Alois Schloegl, and Gary E. Birch, "Comparison of evaluation metrics in classification applications with imbalanced datasets," in *ICMLA*, 2009, pp. 777–782.

[3] Guido Gerig, Matthieu Jomier, and Miranda Chakos, "Valmet: A new validation tool for assessing and improving 3D object segmentation," in *Proceedings of the 4th International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2001, pp. 516–523.

[4] Daniel B. Russakoff, Carlo Tomasi, Torsten Rohlfing, Calvin R. Maurer, and Jr., "Image similarity using mutual information of regions," in *8th European Conference on Computer Vision, ECCV*, 2004, pp. 596–607.

[5] Nguyen Xuan Vinh, Julien Epps, and James Bailey, "Information theoretic measures for clusterings comparison: is a correction for chance necessary?," in *Proceedings of the 26th Annual International Conference on Machine Learning*. 2009, pp. 1073–1080, ACM.

[6] David M. W. Powers, "Evaluation: From precision, recall and F-factor to ROC, informedness, markedness correlation," *Journal of Machine Learning Technologies*, vol. 2, pp. 37–63, 2011.

[7] Chris Buckley and Ellen M. Voorhees, "Evaluating evaluation measure stability," in *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*. 2000, pp. 33–40, ACM.

[8] Halim Benhabiles, Guillaume Lavoue, Jean Phillipe Vandeborre, and Mohamed Daoudi, "A subjective experiment for 3d-mesh segmentation evaluation," in *IEEE International Workshop on Multimedia Signal Processing (MMSP)*, 2010.

[9] Filip Radlinski and Nick Craswell, "Comparing the sensitivity of information retrieval metrics," in *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, 2010.

[10] R. Blanco and H. Zaragoza, "Beware of relatively large but meaningless improvements," Tech. Rep., Yahoo! Research 2011-001, 2011.

[11] Tetsuya Sakai, "Evaluating evaluation metrics based on the bootstrap," in *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*. 2006, pp. 525–532, ACM.

[12] Jin Huang and Charles X. Ling, "Using AUC and accuracy in evaluating learning algorithms," *IEEE Transactions on Knowledge and Data Engineering*, vol. 17, pp. 299–310, 2005.

[13] Enrique Amigo, Julio Gonzalo, Javier Artiles, and Felisa Verdejo, "A comparison of extrinsic clustering evaluation metrics based on formal constraints," *Inf. Retr*, vol. 12, no. 4, pp. 461–486, August 2009.

[14] Nguyen Xuan Vinh, Julien Epps, and James Bailey, "Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance," *J. Mach. Learn. Res.*, vol. 9999, pp. 2837–2854, December 2010.

[15] Luca Busin and Stefano Mizzaro, "Axiometrics: An axiomatic approach to information retrieval effectiveness metrics," in *Proceedings of the 2013 Conference on the Theory of Information Retrieval*, New York, NY, USA, 2013, pp. 8:22–8:29.

[16] Tetsuya Sakai, "On the reliability of information retrieval metrics based on graded relevance," *Information Processing Management*, 2007.

[17] Ellen M. Voorhees and Chris Buckley, "The effect of topic set size on retrieval experiment error," in *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, 2002, pp. 316–323.

[18] H. B. Mann and D. R. Whitney, "On a test of whether one of two random variables is stochastically larger than the other," *The Annals of Mathematical Statistics*, vol. 18, no. 1, pp. 50–60, 1947.

[19] Janez Demsar, "Statistical comparisons of classifiers over multiple data sets," *J. Mach. Learn. Res.*, vol. 17, pp. 30, 2006.

[20] Abdel Aziz Taha, Allan Hanbury, and Oscar Jimenez, "Test data and results of the automatic metric selection method," Tech. Rep., Vienna University of Technology, http://publik.tuwien.ac.at/files/PubDat_229008.pdf, 2014.