Performance Endpoints Discussion

Partha

One thing to keep in mind is that this would be in the regime of a very large number of hypotheses, and one has to get into false discovery rates etc – also the hypotheses are not independent (they are spatially correlated).

Craig

I am still bootstrapping my way into the pathology world, so take what I say with a grain of salt. You mentioned having a binary (pathologist drawn) map of cancer cells to compare with AI algorithms that produce probability maps. It seem like you can generate something like an ROC curve by applying a threshold to the probability map to get a cell-level false positive rate and true positive rate (by comparison with the pathologist drawn map). Then a curve is generated by adjusting threshold. The Wunderlich paper you sent was really about finding optimal algorithms for detection+estimation tasks. It looks to me like you are more in the area of trying to decide precisely what good performance is. In this case I would point you to Adam's earlier paper (I've attached it), where he used utility structures to identify the most appropriate assessment paradigm. The paper described a general utility structure, where special cases lead to ROC, LROC, FROC, etc. A similar kind of thinking may lead you to a defensible (i.e. rational) assessment paradigm for your tasks. I'd be happy to discuss that further with you, if it interests you.

- Wunderlich, A. & Abbey, C. K. (2013), 'Utility as a rationale for choosing observer performance assessment paradigms for detection tasks in medical imaging.' Med Phys, 40, 111903. (LINK)

Partha

Brandon, thanks for the papers. One minor technical comment: in comparing annotations of two reviewers: if the annotations consist of points in space, then one has to compute a metric between two realizations of a point process. This is potentially slightly nontrivial – there is some ad hoc choice involved in deciding when two points are close enough to be considered an agreed upon detect, and when they disagree. Same is true for comparing algorithmic detects with annotater markings. I mention this since we have been grappling with this issue for detects in large mouse brain images with lots of cells … not an issue if the cells/objects are sparse and far apart, but if they start overlapping then ascertaining false detects/rejects becomes ambiguous.

There are different reasonable heuristics of course, but nothing quite as clean as 2-alternative hypothesis testing (in the context of ROC curves).

For comparing region markings – there are standard measures (eg the so called Jaccard similarity metric). We developed a "concordance" metric in the context of comparing different

brain atlases, that had an interpretation as a relative probability, since the Jaccard metric seemed to have some issues to us. Here is the link,

http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0007200

Brandon

Camelyon 16 paper summarizing study design and results

Ehteshami Bejnordi, B.; Veta, M.; Johannes van Diest, P. & et al. (2017), 'Diagnostic Assessment of Deep Learning Algorithms for Detection of Lymph Node Metastases in Women With Breast Cancer.' JAMA, 318, 2199—2210. (LINK)

We need to understand the metrics in this paper and consider alternatives. Jeroen and Mitko have access to the data from AI algorithms and pathologists from the Camelyon16 challenge. The algorithms typically produce a probability map. Pathologists produce individual marks or annotation regions (one pathologist in the Camelyon16 study painstakingly outlined all cancer cells, binary map). I'd like to consider a location based ROC measure of agreement. Camelyon16 uses an FROC method that I would like to understand better, especially in the context of the utility structure (see below) and applications with marks, binary maps, and probability maps.

Location Based ROC methods

We have data from pathologists marking the locations of mitotic figures. I'd like to take two readers and measure the agreement between their marks. Craig Abbey has good experience with ROC generalized to treat data with locations. I know most of these and want to use/formulate an appropriate utility framework for our data/task. Of course, we need to then develop MRMC (multiple-reader, multiple-case) analysis methods for the subsequent methods. MRMC analysis methods account for the variability that comes from the cross-correlated random effects of readers interpreting images. Hopefully, we can use the experience and theoretical foundation of previous work (U-statistics or Generalized Linear Mixed Models).

- Wunderlich, A.; Goossens, B. & Abbey, C. K. (2016), 'Optimal Joint Detection and Estimation That Maximizes ROC-Type Curves.' IEEE Trans Med Imaging, 35, (9), 2164—2173. (LINK)