Janelia farm, conference 'Bioimage informatics', April 5-9, 2008

Translating biomedical image content into useful feature spaces

Nikita V. Orlov, John Delaney, D. Mark Eckley, Lior Shamir, and Ilya G. Goldberg
Laboratory of Genetics, NIA/NIH

Microscopic images of tissue biopsies can be interpreted and sorted efficiently by experienced pathologists, allowing diagnosis of cancerous and non-cancerous states. In theory the same images can be used to train classifiers to generate diagnosis via a machine vision regime. We have developed an approach that translates digital images into a multidimensional feature space, allowing us to search for hidden content in biomedical images. Our unique, two stage approach has leveraged the hidden image content to produce a very accurate and reproducible classifier for two common cancer biopsies – lymph node and melanoma.

We examined a series of schemes (computational chains) for constructing feature spaces, employing a novel two-stage approach. In the first stage, Image filters are applied to an image, producing multi-scale or spectral decompositions. Then a set of global features was computed from the produced decompositions. The resulting feature set was more efficient, because it combines several layers of image content, potentially improving the classification accuracy.

We compared three basic ways to construct image filters: image pyramids, subband pyramids, and image transforms (spectral representations) by classifying a benchmark set of five separate biology problems. We present the performance of several different computational chains applied to the benchmark set. Image transforms produced the most effectual chains, as judged by classification accuracy, for all five problems.

Datasets from two distinct cancer biopsies were classified using the transform-based chains. Lymph node biopsies commonly contain three types of malignancies: chronic lymphatic leukemia (CLL), follicular lymphoma (FL) and mantle cell lymphoma (MCL). We imaged several slides and applied three different classification algorithms. No preprocessing of data was employed, including segmentation, ROI selection or contrast enhancement and yet, the highest classification rate achieved on this set was 0.98.

We also analyzed an image dataset generated from melanoma samples. The data were obtained by imaging stained tissue micro-arrays (TMA) printed on glass slides. Our data contained seven discernable malignant tissue types.

For this set we chose a special way to partition out the test images. We employed the Leave-One-Out approach *on the spot level*: one TMA spot per tissue type was set apart for testing, while other spots combined with all other spots (all tissue types) formed the training partition. Then we rotated spots (for each tissue type) in round robin manner producing four data splits based on different spot sets. This way of splitting data is *medically significant*: not only does the classifier know about test images, it does not even touch the whole spot (as an image pool dedicated for testing). This creates basis for introducing spots with completely different variation of data – like different patients, different disease stage, aggressiveness, etc.

The classification accuracy on the melanoma set was 0.93 (average on all test sets, seven classes).