Pattern Recognition 101

This document is prepared for non-computer scientists. It *loosely* explains some concepts of Pattern Recognition that will be encountered during the use of the tool BIOCAT (BIOimage Classification and Annotation Tool).

What is pattern recognition? -- The ability of a computer to automatically recognize patterns from data, such as images. It typically involves assigning a label to the data sample. We will assume it works with images (or regions in images) in the rest of the document since those are what BIOCAT works with.

What is a typical pattern recognition work flow? -- A pattern recognition module typically starts with extracting a set of features from images such as average intensity of the image (although complex feature go way beyond that), then optionally select some "good" features from the extracted features (see below for feature selector).

Why can pattern recognition help biology? -- a) High-throughput microscopic imaging needs automated approach for image classification and annotation to improve efficiency and reduce bias; b) High-content (often high-dimensional) microscopic images need pattern recognition to improve quantifying objects inside images or modeling dynamics among images; c) Exploring the most useful features for analyzing dis/similarity among different categories of images, which may contribute to scientific discovery.

What is the training/testing mode for model selection? -- Taking two sets of images, one for training, another for testing, both are labeled with *desirable targets*. The training process builds a model is on the training set, and the testing process comes up with *output labels* for each testing image using the model. Different models can be built to choose the one that does the best on the testing set (measure by how the output labels match the desirable target label using recognition rate).

What is the cross-validation mode for model selection? -- Taking one set of images, divide into roughly equal size subsets. Treat every subset as the testing set and the rest as the training set, run the training/testing experiment described above, repeat for each subset, and calculate the average of the recognition rates of all runs. The model is chosen based on the average.

- What is a fold in cross-validation? -- The number of divisions (which is the same as number of runs)
- When do I need shuffling in Cross-validation? -- If the images are ordered based their category, shuffling can void the situation when training images are unrepresentative of the testing images.

What is classification? -- It is the so called "supervised learning" studied in machine learning field, which is the process when an algorithm assigns a label to a data sample after training was conducted. Many classification algorithms exist such as Nearest Neighbor and Support Vector Machine. For a non-computer scientist, this can be understood as the step when the label is assigned.

What is annotation, and how does it relate to classification? -- In the context of "biological image annotation", this is the step that assigns one label or multiple labels to an image (or a region in an image). In BIOCAT, annotation is formulated into a pattern recognition problem, so that a label or multiple labels can be assigned by the classification algorithm.

What is a feature extractor? -- The algorithm that extracts a set of features from an image. Features can be simple statistics such as average intensity of the image pixels, although complex features go way beyond that. BIOCAT provides many feature extractors as candidate modules. An automatic comparison mechanism can be used to find the one or several combined features (in a chain together with the feature selector and classification algorithm, see below) that are most suitable for the task at hand.

What is a feature selector? -- It has been known in pattern recognition field that sometimes a subset of features may work better than the whole set, for example, irrelevant or redundant features may hurt the algorithm. In addition, using a subset also reduce the computational complexity of the entire model. Feature selector selects a "good" subset of features from existing features. The criterion of being "good" features varies in different selecting algorithm. In general, we want the good features to be the most representative of the corresponding category of the image.

What is an algorithm chain? -- It represents one pattern recognition model that typically consists of multiple feature extractors, some (optional) feature selectors and a mandatory classifier.

What are the parameters used by the algorithms? -- Each feature extractor/selector or classifier can be associated with parameters to define some fine-grained behavior. For non-computer scientists, the default setting is usually fine.