Image Classification of White Blood Cells with Few Examples

Vineet Rai

Business Objective

The task of classification, or labeling, of images based on training examples falls under the scope of *supervised machine learning*. A good supervised learning model should learn as much as possible from its sample data without overgeneralizing from it, since this can lead to reduced accuracy on unseen examples, a phenomenon known as *overfitting*. When the number of training examples is low, it is even more critical that an algorithm extracts and learns from all salient features while avoiding overfitting.

What follows is a demonstration using 139 files from the LISC database, a collection of hematological images containing stained white blood cells in peripheral blood. White blood cell classification can be used to diagnose disease, which can elevate or depress certain types of cells in the blood. To simplify the task, only images containing a single lymphocyte, monocyte, or neutrophil were selected.

After splitting the data very unevenly so that only 10% (13 images) are training examples, and other 90% (126 images) are reserved for testing, what is the highest testing accuracy that can be achieved?

Solution Overview

Image Processing Using GNU Octave (MATLAB)

- Preprocessing: color channel, pixel intensity, smoothing, edge detection, object separation
- Segmentation: multilevel thresholding, object labeling, cell nucleus identification
- Feature Extraction: geometric properties of the cell nucleus as numeric attributes

Machine Learning and Classification Using R

- Shuffle Data
- Assign Train/Test Split
- k-Nearest Neighbors Classification Algorithm

Results Obtained

The k-Nearest Neighbors algorithm achieved an accuracy of 92.9% when classifying images from the training set, using just 13 training examples.