Fuzzy Decision Tree for Accurate Breast Cancer Diagnosis

Jed Lim, Karimi Zayan, Matthew Kan, Makuyh Das

12 October, 2022

1 Background

Breast cancer is the number one most common cancer amongst women in Singapore. Early and accurate diagnosis of breast cancer is important for breast-saving and life-saving treatment.

The gold standard for the diagnosis of breast cancer is by surgically removing the breast lump with a complete microscopic examination of the breast tissue to look for cancer cells.

Fine needle aspiration is an alternative that allows the doctor to take out a small amount of tissue from the breast lump, without the need for surgery to remove the entire breast lump. By examining the characteristics of the cells, doctors have been able to diagnose breast cancer with variable success. Increasing the success of fine needle aspiration allows for diagnosis of breast cancer without the need for a woman to undergo surgery to remove the breast lump.

To resolve this, this project uses a fuzzy decision tree to classify breast tumor cells into malignant cancer cells or benign non-cancerous cells.

2 Dataset

The dataset used was the Breast Cancer Wisconsin (Diagnostic) Data Set from the University of Irvine (UCI) Machine Learning Repository.
The dataset contained 569 instances, with no missing data. 357 instances were benign (not cancerous) and 212 were malignant (cancerous).

The features were computed from digitalized images of fine needle aspirates of breast tumors.

The features describe 10 characteristics of the cell nuclei present in the images:

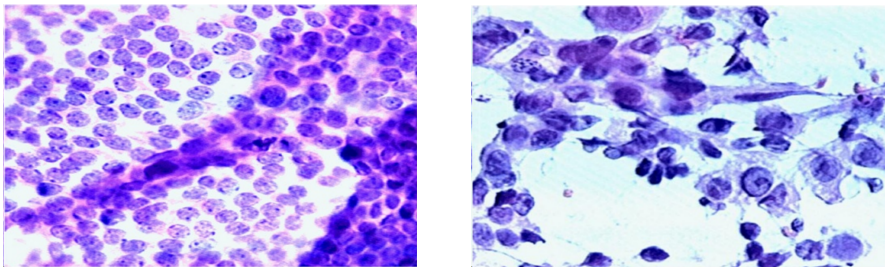

Figure 1: A picture of breast cells. The cells on the left are begin while the cells on the right are cancerous.

- The radius of an individual nucleus
- The nuclear perimeter
- The nuclear area
- Compactness of the nucleus
- The smoothness of the contour of the nucleus
- The number of contour concavities
- The symmetry of the nuclear contour
- The texture of the cell nucleus

The mean, standard error and worst (mean of the three largest values) of these features are computed for each image, resulting in 30 features in the UCI dataset.

We follow in the footsteps of Sizilio et al. **??** and add newly generated features of homogeneity and uniformity that were demonstrated to have diagnostic importance.

Uniformity is the difference between the radius worst value and the radius mean value and is an indication of the variability in size of the cell nuclei.

Homogeneity is the difference between the worst value of symmetry and the mean value of symmetry and is an indication of the symmetry of the cell nuclei.

According to Sizilio et al., the features of area, perimeter, homogeneity and uniformity produced the best results. Thus, we use these 4 feature and drop all other features.

The minimum and maximum area, perimeter, homogeneity and uniformity for the 2 labels (benign and malignant) are computed and displayed in the table

below.

Fuzzy intervals are present for each of the above 4 features, whereby the benign values are within the range of the malignant values. This means that it is not linearly possible to diagnosis a breast lump as benign or malignant using a simple decision tree. Thus, we shall resort to using a fuzzy decision tree.

3 Fuzzy Decision Tree

A typical decision tree works by building a set of if statements to determine which class something belongs to.