An approach to breast tumor classification using an active learning method

Shaan Varia
Carnegie Mellon University

Akul Penugonda
Carnegie Mellon University

December 10, 2014

1 Introduction

We want to model the severity of breast cancer tumors based on factors such as a BI-RADS assessment, age, shape of mass, margin of mass, and mass density. BI-RADS stands for ”Breast Imaging-Reporting and Data System”, and is a categorical assessment of tumor malignancy from 1-5, 5 being the most severe. Shape of mass is the qualitative shape of the tumor, margin of mass is a categorical assessment of the edge region of the tumor, and mass density is simply how the density of the area is distributed. Severity is an indicator of whether the tumor is benign or malignant. These tumors are initially screened for with a mammogram, which provides all of these data besides the severity. A biopsy, which is rather expensive, is necessary to get the severity. It would be great to have a learner that can predict whether a tumor is benign or malignant without a biopsy.

2 Background

2.1 Related Work

3 Methods

3.1 Data Source

We will get our data from the UCI machine learning repository. The below data set has 961 entries, where each entry contains the following features and labellings:

6 Attributes in total (1 goal field, 1 non-predictive, 4 predictive attributes)

1. BI-RADS assessment: 1 to 5 (ordinal, non-predictive!) 2. Age: patient's age in years (integer) 3. Shape: mass shape: round=1 oval=2 lobular=3 irregular=4 (nominal) 4. Margin: mass margin: circumscribed=1 microlobulated=2 obscured=3 ill-defined=4 spiculated=5 (nominal) 5. Density: mass density high=1 iso=2 low=3 fat-containing=4 (ordinal) 6. Severity: benign=0 or malignant=1 (binominal, goal field!)

Taken from https://archive.ics.uci.edu/ml/datasets/Mammographic+Mass

We will not use BI-RADS as it is too predictive of a feature, we will instead just use Age, Shape, Margin and Density to predict Severity.

3.2 Features

We will write an active learner which looks at four features obtained from a mammogram of the breast. These four features are age of patient, shape of breast, margin of breast, and density of breast. We will use these to predict severity, i.e. whether the mass in the patients breast is benign or malignant. Before discussing the learner, we will go over the role that each of these features play in relation to the mass observed being malignant or benign.

- Density

  Density of mammographic mass does not refer to the traditional definition of density (mass/volume), rather it refers to the amount of fat and tissue in the breast. A dense breast is one that has more tissue than fat. Studies have shown that women with high breast density are 4-5 times more likely to develop breast cancer compared to women with low breast density [1].

Density is an ordinal field, that is, the higher the density the higher the chance of the mass being malignant.

- Margin

  Margin refers to what the outer part of the mass looks like. Margin of the mammographic mass is a nominal field, and as such we will go into the distinctions between the classes.

  - Circumscribed

    The mass is surrounded by tissue and is not as likely to be malignant, most circumscribed margins are likely to be benign [2,3].

  - Microlobulated

    Microlobulated breast margins refer to many small lobulations on the surface of a breast nodule. If more of these microlobulations pop up on the mass then there is a higher chance of the mass being malignant [4], and as such microlobulated masses are treated with more care, and are not simply passed off as benign.

  - Obscured

    Obscured breast margins are suspicious due to the fact that they cannot be discerned. Ill-defined/Irregular An ill defined margin is one that cannot be distinguished from the fatty tissue surrounding it. This may mean that the mass is invading the tissue, which would be an indicator of malignancy.

- Shape

  Shape is similarly a nominal field, and simply refers to the visual shape of the mass as observed on the mammogram. Round, Oval, Lobular All of these shapes are not highly indicative of malignancy. Lobular masses refer to those who have small undulations on the mammogram. Irregular Irregular masses are indicative of malignancy by virtue of the fact that they are not the regular shapes of common benign masses such as fibroadenomas or cysts.

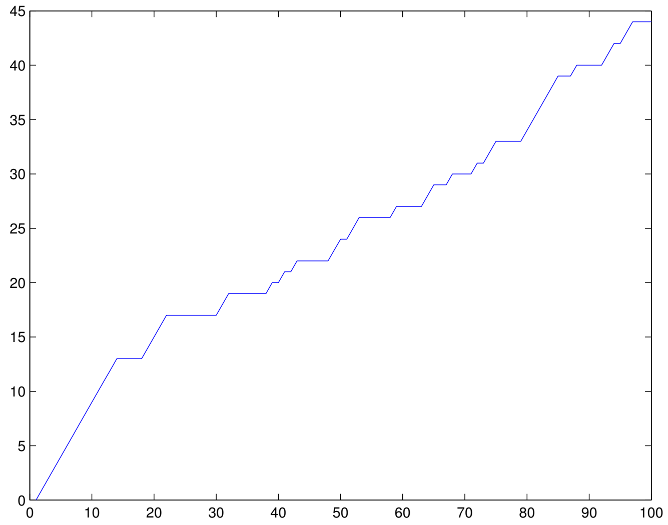

Figure 1: Cost Curve

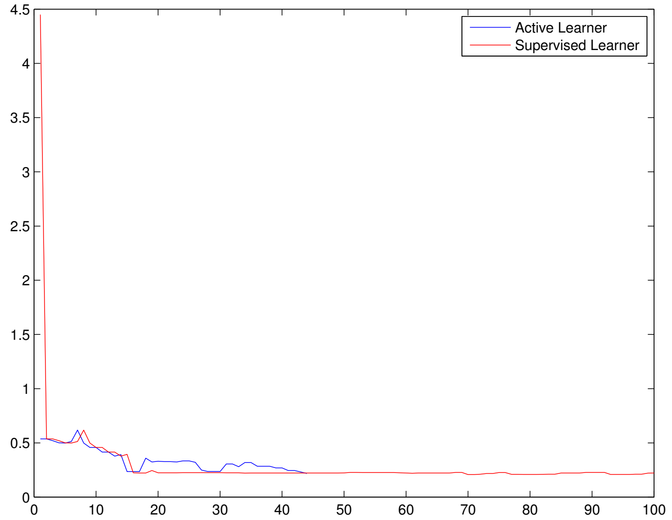

Figure 2: Generalized Error

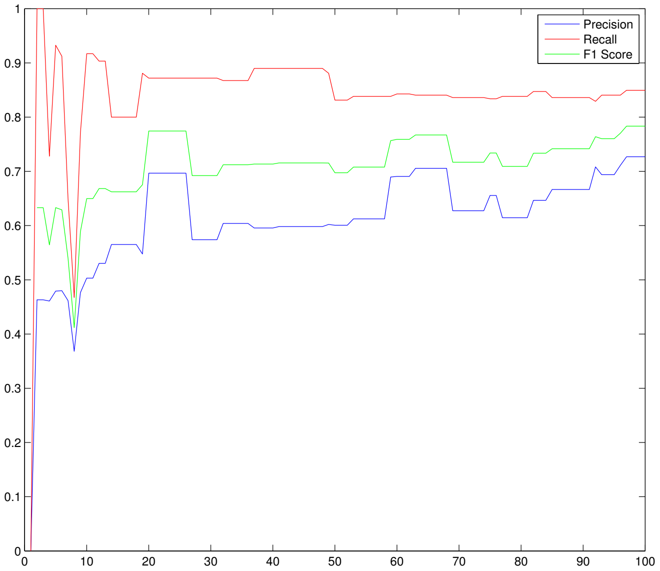

Figure 3: Generalized Error

3.3 Active Learning

4 Results

5 Discussion

6 Results

6.1 Future Work