Wisconsin Breast Cancer Diagnosis Deep Learning Revisited

(Capstone Proposal)

By

Yuefeng Zhang

1. Domain Background

The machine learning methodology has long been used in medical diagnosis [1]. The Wisconsin Diagnostic Breast Cancer (WDBC) dataset [1] has been widely used in research experiments.

Most of publications focused on traditional machine learning methods such as decision trees and decision tree-based ensemble methods [5].

Recently supervised deep learning method starts to get attention. For instance, Stahl [3] and Geekette [4] applied this method to the WDBC dataset [1] for breast cancer diagnosis using feature values calculated from digitized image of a Fine Needle Aspirate (FNA) of a breast mass.

Geekette [4] used the WDBC dataset with derived features (e.g., mean, standard error, …, etc.) and experimented three types of deep neuron networks: 1, 2, and 3 hidden layers of 30 neurons without any data pre-processing. Stahl [3] used only the originally identified features with pre-processing such as center and scale, but did not provide the details of the neuron network architecture in use.

In this proposal, I propose a new supervised deep learning method for analyzing the same WDBC dataset for breast cancer diagnosis.

2. Problem Statement

The FNA of a breast mass has been used in advancing breast cancer diagnosis [1][2]. To this end, features are computed from such digitized image. These features describe the characteristics of the cell nuclei present in the image.

Problem:

Given the list of feature values calculated from a digitized image of the FNA of a breast mass from a patient, how to diagnose (determine) whether or not the patient has breast cancel?

This problem is treated as a 2-class classification problem (*benign* or *malignant*).

3. Datasets and Inputs

The WDBC dataset is publically available and can be downloaded at: https://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/breast-cancer-wisconsin.data

As described in the WDBC dataset [1], the following features were computed for each cell nucleus and will be used as inputs:

- Clump Thickness
- Uniformity of Cell Size
- Uniformity of Cell Shape
- Marginal Adhesion
- Single Epithelial Cell Size
- Bare Nuclei
- Bland Chromatin
- Normal Nucleoli
- Mitoses

All of the 699 samples in the dataset will be used in the proposed deep learning method. Similarly to [4], this dataset will be split into two parts: 75% for training and the remaining 25% for testing.

The last column "Class" in the dataset [2] will be used as label in training the new proposed deep learning network.

Data pre-processing for balancing classes will be applied on a needed basis according to the result of implementation.

4. Solution Statement

In this proposal, I propose a new supervised deep learning method to inherit the merits of the methods experimented by Stahl [3] and Geekette [4].

Specifically, similarly to [4], the new method uses the originally identified features [2], pre-processes data using center and scale, and treats the problem as a classification problem. Like [3], the new method adopts a three hidden layers of deep neuron network (see Project Design section for details).

5. Benchmark Model

The result of the new supervisor deep learning method will be compared with the result of applying the out-of-box implementation of the sk-learn Random Forest Classifier [6] to measure the relative performance.

6. Evaluation Metrics

Similarly to the evaluation method practiced by Geekette [4], The AUC will be used as the evaluation metrics and F1 will be considered as an alternative.

7. Project Design

The architecture of the new deep neuron network is shown in the following figure:

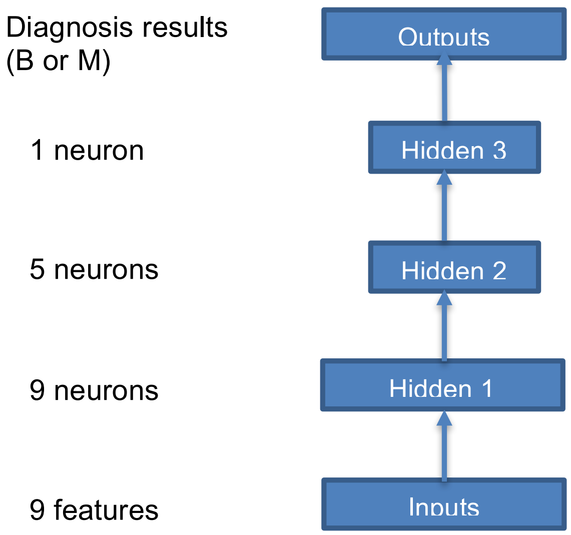

Figure 1: The architecture of the new deep neuron network.

The inputs to the network are the feature values computed from digitized image of FNA, and the outputs from the network are the two identified classification classes B (*benign*) or M (*malignant*).

I select the first hidden layer of 9 neurons and third hidden layer of 1 neuron to match the number of input features and the number of output classes respectively. The size of the second hidden layer is in the middle of the sizes of the first and third hidden layers.

References

[1] W. H. Wolberg, etc., “Wisconsin Diagnostic Breast Cancer (WDBC)”: https://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/breast-cancer-wisconsin.names , November 1995

[2] WDBC data file: https://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/breast-cancer-wisconsin.data , February 5, 1996

[3] K. Stahl, “Wisconsin Breast Cancer Diagnosis Deep Learning”:

http://www.rpubs.com/kstahl/wdbc_ann, July 17, 2017

[4] D. Geekette, “Breast Cancer data: Machine Learning & Analysis”:

http://rpubs.com/elena_petrova/breastcancer, November 25, 2016

[5] L. a comment, “Random forest classifier combined with feature selection for breast cancer diagnosis and prognostic”, Scientific Research, Vol.6 No.5, May 2013, http://www.scirp.org/journal/PaperInformation.aspx?PaperID=31887

[6] sk-learn Random Forest Classifier: http://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html