Wisconsin Breast Cancer Diagnosis Autoencoder

(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]. Both of these practices relied upon human labeled answers in training data. So far published results are rare in applying self-supervised or unsupervised deep learning to WDBC dataset.

In this proposal, I propose a new self-supervised autoencoder [6] deep learning method for analyzing the same WDBC dataset for breast diagnosis.

2. Problem Statement

Digitized image of a Fine Needle Aspirate (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.

As described in the WDBC dataset [1], the following ten features are computed for each cell nucleus:

- radius (mean of distances from center to points on the perimeter)
- texture (standard deviation of gray-scale values)
- perimeter
- area
- smoothness (local variation in radius lengths)
- compactness (perimeter^2 / area - 1.0)
- concavity (severity of concave portions of the contour)

- concave points (number of concave portions of the contour)
- symmetry
- fractal dimension ("coastline approximation" - 1)

Problem:

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

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/wdbc.data

4. Solution Statement

In this proposal, I propose a new autoencoder [6] method to explore the feasibility of applying self-supervised deep learning to WDBC dataset for breast cancel diagnosis.

5. Benchmark Model

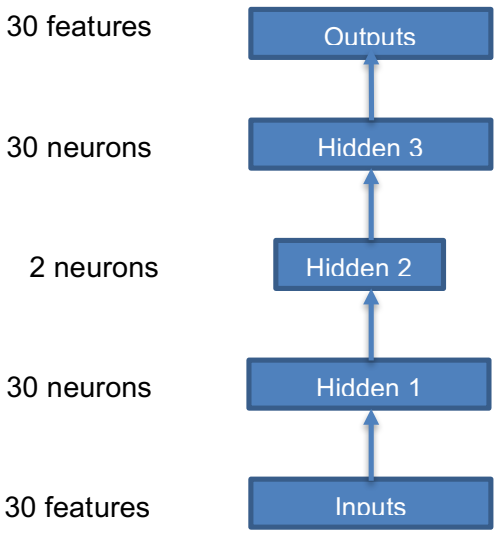
The result of the new autoencoder deep learning method will be compared with the result of the supervised deep learning method by Stahl [3] to measure the relative performance.

6. Evaluation Metrics

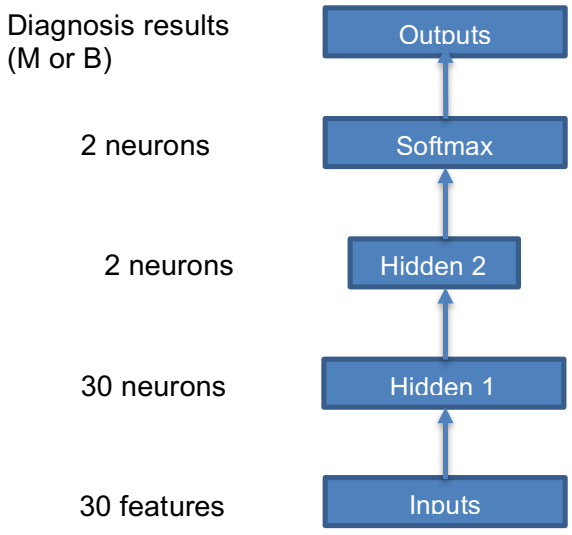
Similar to the evaluation method practiced by Geekette [4], Both of AUC and total accuracy rate will be used as evaluation metrics.

7. Project Design

The architecture of the autoencoder and autodecoder is shown in the following figure:



Once the above autoencoder and autodecoder have been trained, the trained autoencoder will be combined with a softmax layer to form the target deep learning network as follows for breast cancer diagnosis:



References

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

[2] Index of WDBC: https://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/, 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] F. Chollet, "Building Autoencoders in Keras": https://blog.keras.io/building-autoencoders-in-keras.html, May 14, 2016.