Wisconsin Breast Cancer Diagnosis Deep Learning Revisited

By

Yuefeng Zhang

1. Introduction

The machine learning methodology has long been used in medical diagnosis [1]. The Wisconsin Breast Cancer Database (WBCD) dataset [2] 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 WBCD dataset [2] for breast cancer diagnosis using feature values calculated from digitized image of a Fine Needle Aspirate (FNA) of a breast mass.
These features describe the characteristics of the cell nuclei present in the image.

Given the list of features calculated from a digitized image of the FNA of a breast mass from a patient, the problem is how to diagnose (determine) whether or not the patient has breast cancer. This problem can be treated as a 2-class (*benign* or *malignant*) classification problem.

Stahl [3] used the WBCD 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. Geekette [4] 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.

This capstone project report presents a new supervised deep learning method for analyzing the same WBCD dataset [2] for breast cancer diagnosis using common open source libraries. The new supervised deep learning method inherits 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 2-class classification problem. Like [3], the new method adopts a three hidden layers of deep neuron network.

The rest of this report consists of 5 sections. Section 2 discusses the dataset, and related issues and pre-processing solutions. Then Section 3 describes the new supervised deep learning method. After that, Section 4 discusses the performance of the new method. Finally Section 5 is the conclusion.

2. Dataset and Dataset Pre-Processing

The dataset used in this project is the publically available WBCD dataset [2]: https://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/breast-cancer-wisconsin.data

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

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

The following issues were observed with the dataset:

- Missing data
- Small dataset size
- Various ranges of data values
- Skewed data

2.1 Missing Data

One of the common issues with dataset is missing data and there is no exception to the WBCD dataset. Specifically quite a few Bare Nuclei entries are missing (marked as ?). There missing data entries are replaced with 0 for simplicity in this project.

2.2 Small Dataset Size

The original WBCD dataset [2] contains only 699 samples, which is too small for deep learning after dividing the dataset into training and testing subsets.

This data size issue is resolved in this project by generating new data samples as follows. First the *Numpy.random.normal*(*mean, sigma, features*.*shape*) method is used to generate an array of the same dimensions as the original dataset with normal distributed random numbers, where *mean* = 0 and the standard deviation *sigma* = 0.1. Then the generated array of random numbers are added into the original dataset element by element to form a new dataset.

The last column "Class" in the WBCD dataset [2] is used as label in training deep learning model. These labels (i.e., the values of the Diagnosis or Class column) are used to label the corresponding samples (feature vectors) in the generated dataset. This data generation process is repeated to generate multiple new datasets. Finally all of the new datasets are combined with the original dataset to form a final dataset with 30,756 samples in total for experiment in this project.

2.3 Various Ranges of Data Values

The ranges of feature values for different features are different. As a common practice, for each of the features, the feature values are scaled into the range of [0, 1] as follows for deep learning:

(*Value* – *Minimum*) / (*Maximum* – *Minimum*)

2.4 Skewed Data

Figure 1 shows the Pandas generated scatter plot matrix of the dataset after data pre-processing as described in Sections 2.1, 2.2, and 2.3. It can be seen from the figure that the dataset is skewed to the right.

This data skewness issue is handled by taking squired root of the feature values. Figure 2 shows the result of applying squired root transformation to the dataset.

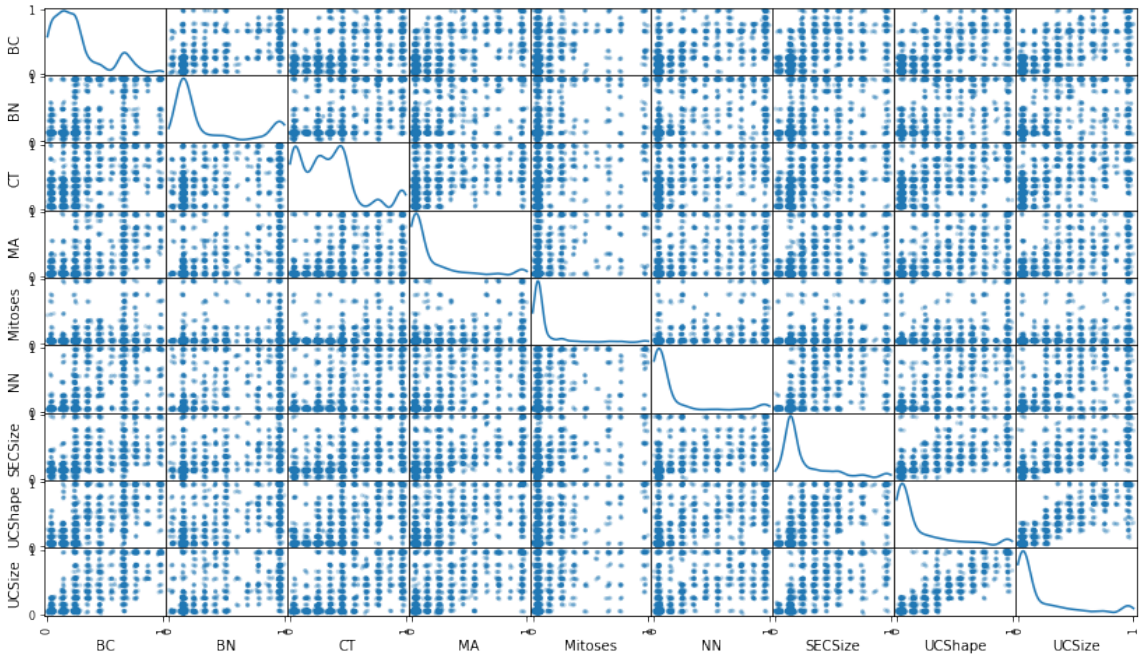

Figure 1: Skewed dataset.

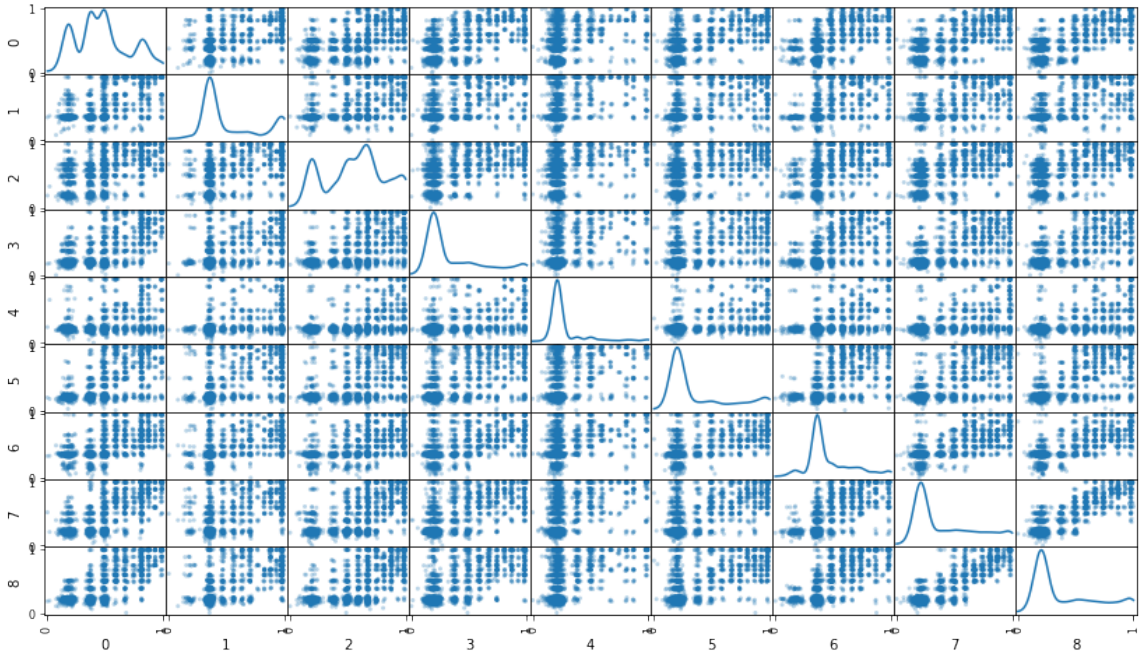

Figure 2: Transformed dataset.

3. New Deep Learning Method

As shown in Figure 3, similarly to [3], four hidden layers are used in the new deep learning network. In addition, dropout layers with dropout rate of 0.5% are used to avoid overfitting and improve network performance in accuracy.

The common activation function *relu* is used in hidden layers 1, 2, and 3. The *sigmoid* activation function is used in the final hidden layer to produce continuous output in the range of (0, 1).

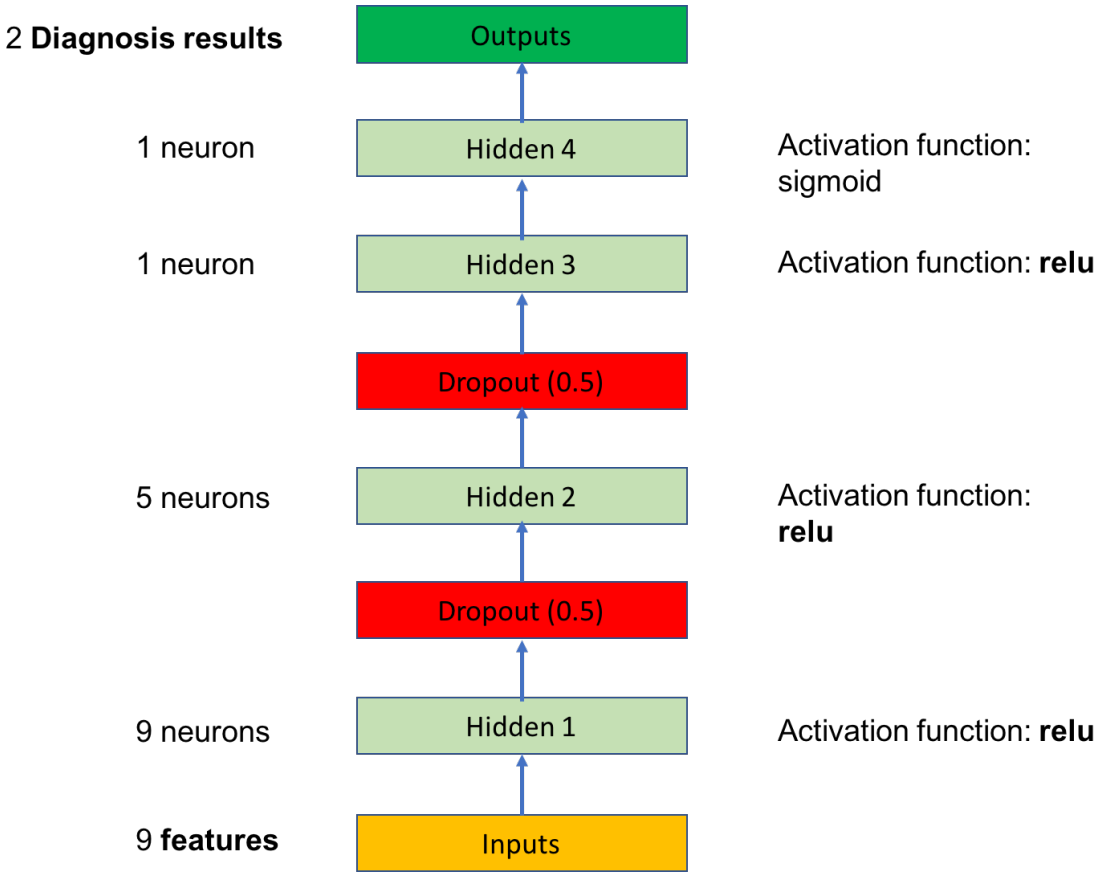

Figure 3: The architecture of the new deep neuron network.

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

The first hidden layer uses 9 neurons and the last two hidden layers use 1 neuron to match the number of input features and the number of output classes respectively. The size (5 neurons) of the hidden layer in the middle is in the middle of the sizes of the first and last two hidden layers.

The popular open source Keras deep learning library [7] is used for implementation in this project.

Similarly to [4], the dataset of 30,756 samples is split into two parts: 75% for training and the remaining 25% for testing.

The implementation details of the project are available in GitHub [8].

4. Results

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.

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.

5. Conclusion

References

[1] W. H. Wolberg, etc., "Wisconsin Breast Cancer Database (WBCD)": https://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/breast-cancer-wisconsin.names , January 8, 1991

[2] WBCD data file: http://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/breast-cancer-wisconsin.data, January 8, 1991

[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

[7] Keras: The Python deep learning library: https://keras.io/

[8] Y. Zhang, Capstone Project in GitHub https://github.com/mlyuefeng/machine-learning/tree/master/capstone