Report

Yuqi Huang

Method and Procedure

1. Problem formulation and notation

- Instances X = {xi}, where i = 1,2,…,100. 100 is the number of images inside the bag. Here in this simpler version of MIL model, X is similar with a bag of patches from the sample's slides in the paper.
- A bag label Y is the result of fraction x, which represents the percentage of digit 0 in a bag. Here in this simpler version of MIL model, Y is similar with the percentage of cancer cells within the tumor sample in the paper, which is called tumor purity.
- The objective is to predict Y from a given bag of X.

2. Data collection

The MNIST data was obtained from Kaggle: https://www.kaggle.com/oddrationale/mnist-in-csv. A bag was generated with 100 images, containing only digit 0 and digit 7. This bad was used as the training data set.

3. Train a model using the bag of images

Stage One: Feature Extraction and Pooling filter

Use random forest to select the important features and construct the feature space. Boruta algorithm, which uses Random Forest, was conducted to select features. The method performs a top-down search for relevant features by comparing original attributes' importance with importance achievable at random, estimated using their permuted copies, and progressively eliminating irrelevant features to stabilize that test (R Core Team, 2019).

Stage Two: Bag-level representation transformation

Model was trained, and cross validation was done on the training data(a bag with 100 images). The whole date set was split randomly into 10 parts, with 10 samples each time in the validation set. In 10 iterations, 1 of the 10 parts took turn to be validation set, and the other 9 parts became training set for cross-validation.

For the neural network model, the number of layers was set to be three. The parameter of "hidden" was adjusted in the process. "hidden" is a value allows us to specify the number of perceptrons in the hidden layer. The value of "hidden" was set to be 1 to 10. For each "hidden", the square residual of validation set in cross-validation was calculated.

4. Test the model

After choosing the number of perceptrons in the hidden layer, this result neural network model was applied to test dataset.

10 test bags were generated from the original data, each with 100 images. The result NN model was used to predict Y (the percentage of digit 0 in a bag) for each test bag respectively. The model is evaluated by calculating the residual sum of square of all test bags.

Results

Feature extraction:

Boruta output important features. 79 features were extracted for each image inside the bag, so I obtained a feature matrix (dim = 79 x 100) constructed from extracted feature vectors.

```
Boruta performed 499 iterations in 1.635079 mins.
 77 attributes confirmed important: X10x17, X11x17,
X12x17, X12x18, X13x12 and 72
more;
 687 attributes confirmed unimportant: X10x1, X10x10,
X10x11, X10x12, X10x13 and 682
more;
 20 tentative attributes left: X12x19, X13x23, X14x22,
X14x23, X15x9 and 15 more;
```

Bag-level representation transformation

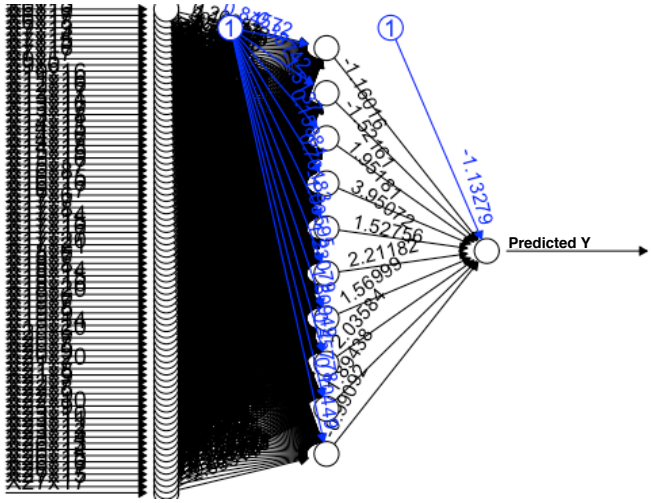
When the number of perceptrons in hidden layer ranges from 1:10, the errors(residual square) are: 26.35849, 26.15668, 25.85168, 25.63518, 25.53549, 25.41643, 24.73684, 24.56193, 24.44193, 24.42478. Therefore, I chose the total number of perceptrons in hidden layer as 10, because its error in cross validation is the smallest.

Testing result:

After using the 10 test bags to evaluate the model, the residual sum of square (RSS = ∑(true_Y - predicted_Y)^2) was obtained, with the value of 16.39125.

Visualization result:

The following is the NN model trained above. The number indicates the weight for each node.



Discussion

This simpler model showed the basic process of MIL model in the paper. Firstly, MIL model uses a given bag of X, which is a collection of patches over a sample's slides, as the predictor. In my model, I used a given bag of X, which is a collection of images containing only 0 and 7, as the predictor. Secondly, MIL model uses the genomic tumor purity of a sample as the response Y. In my model, I used the percentage of digit 0 in a bag as the response Y. Thirdly, in this model, I used three-layer multilayer-perceptron neural network model, which is similar with what has been done in the MIL model.

Some differences exist between this simpler model and the MIL model. In the paper, the MIL model conducts feature extraction by neural networks, and it conducts pooling filter by estimating a marginal distribution over each extracted feature (Oner et al., 2021). In the simpler model I implemented, I used random forest as the feature selection method.

Besides, this simpler model will show the MIL model better if the dataset contained images representing the top and bottom slides in the tumor sample, for the reason that predicting a sample's tumor purity using both the top and bottom slides together is better than using only one of them whenever possible.

References

Oner, M., Chen, J., Revkov, E., James, A., Heng, S. Y., Kaya, A., . . . Lee, H. (2021). *Obtaining Spatially Resolved Tumor Purity Maps Using Deep Multiple Instance Learning In A Pan-cancer Study*.

R Core Team (2019). R: A language and environment for statistical computing. R Foundation for. Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.