Classification Algorithm Application in SARS Diagnosis

Chen Yang u6905855

Research School of Computer Science
The Australian National University
Canberra, Australia
clyde.yang@anu.edu.au

Abstract. Computer algorithms are used separately in the medical field nowadays and some diseases could be diagnosed well by inputting related pathological signs data. In the paper, artificial neural network (ANN) and fuzzy signature are used. Firstly, an ANN is created to distinguish SARS from normal people and other patients in this work and using the distinctiveness method to prune the hidden layer. We conclude that the distinctiveness method is effective to prune redundant neurons for a trained neural network with a small number of neurons and keeps the high accuracy and low loss. Then, the fuzzy signature is used to analyse the same data, and the fuzzy signature is also suitable for this problem.

Keywords: artificial neural network, distinctiveness pruning, medical diagnose, fuzzy signature, classification

1 Introduction

Patients with the same disease are believed to have related pathological signs which are the most evidence for doctors to diagnose. However, not all disease symptoms are specific to only one disease and often the symptoms are overlapping [1].

SARS, severe acute respiratory syndrome, is a special kind of pneumonia. Patients with SARS usually begin with high fever (38°C or above), sometimes associated with chills, rigours, headache, malaise, muscle pain, or even diarrhea [2]. In this paper, algorithms are trained to distinguish SARS, high BP, pneumonia patients, and normal people.

As an artificial neural network (ANN) could handle numerous data and model nonlinear systems with a complex system, ANN is also inferred as a powerful tool in medical diagnostics and clinical management [1]. This classification task is easy for ANN to reach 100% accuracy with a big size of hidden layers and hundreds of epochs. But it would be time-consuming and cost more resources to calcite. To reduce the size of the hidden layer, Gedeon [3] created a distinctiveness measure to prune the hidden layer in progressive image compressive, which inspired me to prune my ANN by a similar method.

The fuzzy signature is suitable for this SARS data. The pathological signs have been processed to fuzzy signals, such as "slight", "moderate", and "high" for fever. Also, these signatures contain three levels of hierarchies, which can be aggregated using different aggregations.[2]

I would start with 30 hidden neurons deliberately and try distinctiveness measures to prune neurons while keeping the accuracy and loss and find a small hidden layer artificial neural network. Then, use the fuzzy signature to analyse the SARS data to compare with the ANN method.

2 Methodology

2.1 Data Preparation

4 pre-processed datasets including pathological data for SARS, high BP, pneumonia patients, and normal people from [4] are used in this study. Each dataset has 1000 rows and 23 columns of data from 1000 people. Fever, blood pressure, conditions of nausea, and abnormal pains are important symptoms for potential diseases, especially SARS. Each symptom has been divided into a group of fuzzy sets as follows [4].

- Fever: Slight, Moderate, High
- Blood Pressure Systolic/Diastolic: Low, Normal, High
- Nausea: Slight, Medium, High

- Abnormal Pain: Yes, No

Doctors know that for certain symptoms, such as SARS and pneumonia, they need to check the patient for possible fever, blood pressure, conditions of nausea, and abnormal pains. Moreover, fever needs to be monitored four times a day. And each symptom check has been divided into some fuzzy sets, such as "slight", "moderate", and "high" for fever, "low", "normal", and "high" for both blood pressure types, "slight", "medium", and "high" for nausea, and "yes", and "no" for abnormal pain. And the row data has been normalized to (0,1). Also, the signature has three layers of hierarchies, which can be aggregated by different aggregations.

There is no disease label of people in the original datasets, thus a column is added to record the disease or health at last as the target output of ANN for supervised training. Numbers are used to representing the status of the diseases as follows.

- 0: Normal
- 1: SARS
- 2: Pneumonia
- 3: High BP

Merge these 4000 rows data from 4 csv files to one csv file for ANN. Input the dataset to the ANN system and split it into 80% for training and 20% for test. While, for fuzzy signature, we save the merged table as one txt file without the last column and one txt file only includes the last column.

2.2 Artificial Neural Network Structure

2.2.1 Artificial Neural Network Structure

Firstly, a most simple one-layer fully connected feedforward neural network was built using PyTorch, including 23 input attributes in the input layer, 30 neurons with sigmoid active function in one hidden layer, and 4 neurons are in outputs. Each neuron in the output layer represents a target class, normal, SARS, pneumonia, and high BP.

Hidden neurons trained by backpropagation and cross loss, the learning rate starts from 0.01 and the number of epochs is set to 500 initially. Classic SGD is used for the optimizer.

2.2.2 Distinctiveness Pruning

The neuron output activation vector of hidden neurons is determined as the distinctiveness and the cosine similarity of two vectors is the key value of pruning [4]. Firstly, normalize vectors to the range 0 to 1, and the angle between vectors is also normalized to -0.5,0.5. Thus, calculate the cosine similarity, the angle gotten is between 0º to 180º. If the angle between any pair of vectors is less than 15º, these two vectors are close, and neurons may be very similar. One of the similar neurons could be pruned and add its weight to the rest. If the angle between vectors is larger than 165º, which means these two vectors are almost complementary, then these two neurons could also be complementary neurons, which could be pruned. [4]

2.2.3 Evaluation

The ANN model with 30 hidden neurons could easily reach 100% accuracy in 500 epochs. Then the hidden neurons Could be pruned by the distinctiveness method and checked the accuracy. As this is a medical dataset, accuracy would be the most important thing. To keep the high accuracy and evaluate the algorithm clearer, I only prune the pairs that would not affect the accuracy. If pruning decreases the accuracy, the pruned neurons would be put back to the hidden layer. After pruning, the loss and hidden layer size could be used to measure the performance of distinctiveness pruning.

2.3 Fuzzy Signature

2.3.1 Polymorphic Fuzzy Signature Structure

As Mendis [2] mentions, an initial tree is built with 3 levels of hierarchies, which is illustrated in Figure 1. Every leaf represents a column of data and the leaves from left to right are ordered the same with the dataset columns. And each symptom includes several fuzzy sets, such as "slight", "moderate", and "high" for fever. Leaves have initial weights of 0.1 at the beginning, which need to be trained.

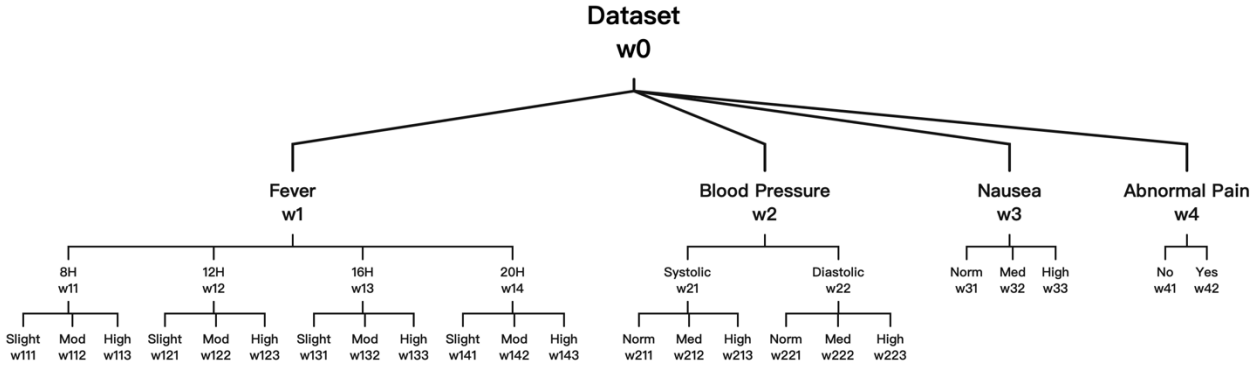

Figure 1 Polymorphic Fuzzy signature of SARS database

2.3.2 WRAO

The weighted relevance aggregation method (WRAO) [6] is used for optimising fuzzy signature. Because compared with others at the same level, some branches may impact more significant to the result. For example, for the fever, the impaction of slight, moderate, and high are less, somewhat, and more. In this case, w111, w112 and w113 will embody this difference.

The WRAO could be written as (1), and p is the aggregation factor. And the sum of squared errors (SSE) is used as the minimise function, which could be expressed as (2).

$$a_{q\ldots i} = \left[\frac{1}{n}\sum_{j=1}^{n}\left(w_{q\ldots ij}a_{q\ldots ij}\right)^{p_{q\ldots i}}\right]^{\frac{1}{p_{q\ldots i}}} \tag{1}$$

where $p \in \Re, p \neq 0, i \in [1,n]$ *and* $\sum_{i=1}^{n} w_i$ *is not necessarily equal to* 1.

$$f(x) = \frac{1}{2}\sum_{i=1}^{m} r_i(x)^2 = \frac{1}{2}\|r(x)\|_2^2 = r^T r \tag{2}$$

3 Results and Discussion

3.1 Performance of the ANN before pruning

Table 1. The loss and accuracy of a simple neural network with 30 hidden neurons during training 500 epochs

Epoch	Loss	Accuracy
1	1.3918	25.31%
101	1.3194	87.13%
201	1.2627	86.81%
301	1.2019	100.00%
401	1.1354	100.00%
500	1.0638	100.00%

Testing Accuracy	100%

Table 1 clearly shows that the simple neural network with 30 hidden neurons could get 100% training accuracy within 300 epochs, and after 500 times train, this network gets 1.2138 for loss and 100% testing accuracy. We may infer that the 30 hidden neurons are redundant, and it is possible to be pruned by the distinctiveness method.

3.2 Performance of the ANN pruning

Table 2. Testing Result of Pruning Neurons by Distinctiveness Method

Neurons Pair	Angle	Type	Loss	Accuracy
6, 8	6.9243º	Similar	1.0687	100.00%
5, 27	8.1893º	Similar	1.0748	78.38%
24, 25	170.6729º	Complementary	1.0975	100.00%
11, 15	11.3412º	Similar	1.0949	100.00%
11, 28	14.5060	Similar	1.0940	92.55%
16, 17	165.2484º	Complementary	1.1170	100.00%
Pruned Neurons	8, 24, 25, 15, 16, 17			

Finding the vectors with the angle that most close to 180º or 0º and the first pair found are No.6 and No.8 neuron. The angle between these two neurons' sigmoid active function is 6.9243º, which means these two neurons are similar. Add No.8 neuron's weight to No.6's weight and No.8 would be pruned by setting the weight to 0. Testing after pruning these two neurons, the loss slightly increases to 1.0687, while the accuracy remains 100.00%.

Continually pruning, the next pair of vectors with an 8.1893º angle are also similar neurons. However, after pruning, the accuracy decreases to 78.38%, therefore this pruning is not as good as the previous one and pruned neurons are put back.

Then a pair of complementary vectors are found, No.24 and No.25, having a 170.6729º angle calculated by cosine similarity. Both neurons could be pruned, and the testing accuracy is still 100.00% after pruning, but the loss goes up to 1.0975.

In this way, compare all pairs of neurons and 6 neurons are pruned. The hidden layer size is reduced by 20%, from 30 to 24, while the accuracy remains at 100% with a tiny increase of loss. More neurons could be pruned if sacrifice some accuracy.

3.3 Performance of Fuzzy Signature

For this dataset, the first 1000 rows are from normal people, which are represented as 0, and the following 3*1000 rows are SARS, Pneumonia and High BP separately, recording as 1, 2 and 3 in the dataset. As Figure 2 shows, the fuzzy signature could generate the result of 4 groups of people.

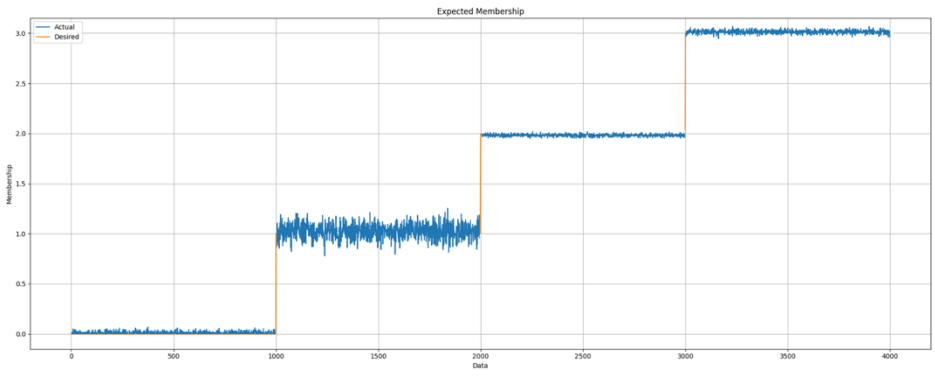

Figure 2 Classification Result by the fuzzy signature method

3.4 Compare and discuss

Table 3. The loss and accuracy of a simple neural network with 24 hidden neurons during training 500 epochs

Epoch	Loss	Accuracy
1	1.4212	24.37%
101	1.3367	74.85%
201	1.2970	85.30%
301	1.2533	100.00%
401	1.2027	100.00%
500	1.1445	100.00%
Testing Accuracy		100%

Table 4. Testing Result of Pruning Neurons by Distinctiveness Method

Neurons Pair	Angle	Type	Loss	Accuracy
1, 17	8.5076º	Similar	1.1510	74.88%
0, 9	14.6693º	Similar	1.1598	100.00%
Pruned Neurons	9			

After pruning by the distinctiveness method, we get a one-layer neural network with 24 hidden neurons. Then I try to set the number of hidden neurons to 24 directly in the code and I get a decent neural network as Table 3 and 4. With only 24 neurons, the neural network could also be trained well, and distinctiveness pruning is still useful to drop a redundant neuron.

Furthermore, the number of neurons is reduced to 10 and set the learning rate to 0.2, we can still get a high accuracy neural network as in Table 5 and 6. When the number of neurons gets small, it would be hard to randomly reduce the hidden layer size, and the accuracy and loss are easy to be affected. In these circumstances, distinctiveness pruning that accurately drops redundant neurons has more advantages.

Thus, distinctiveness pruning could only prune some similar or complementary neurons of a trained network but could hardly find the minimum size of the hidden layer and the necessary number of hidden neurons. It should be more useful in neuron limited questions and smaller size hidden layers. For this dataset, it could be better to reduce the hidden-layer size randomly by try to set the number of neurons lower at first, and use distinctiveness pruning at last to press a surplus zone.

Table 5 The loss and accuracy of a simple neural network with 10 hidden neurons during training 500 epochs

Epoch	Loss	Accuracy
1	1.3923	24.60%
101	1.3297	49.58%
201	1.2765	74.74%
301	1.2013	74.74%
401	1.1003	100.00%
500	0.9800	100.00%
Testing Accuracy		100%

Table 6 Testing Result of Pruning Neurons by Distinctiveness Method

Neurons Pair	Angle	Type	Loss	Accuracy
2, 9	173.3979º	Complementary	1.0763	100.00%
Pruned Neurons	2, 9			

We used polymorphic fuzzy signature structure and WRAO method to optimise the structure in this paper. Compared with Wong's straightforward method [6], which is proved to be stable and performs better than Kóczy's methods [4], the polymorphic fuzzy signature structure uses fuzzy constraints at laves and every leaf has its independence weight.

In the further experiment, we found that only 6 neurons are enough for this dataset to have 100.00% accuracy. Although the fuzzy signature is suitable for this question, it costs a little bit more time than an ANN method for this database. As this dataset has almost linear regression between symptom data and disease result, these methods are all able to distinguish 4 classes well.

4 Conclusion and Future Work

The destituteness pruning method is demonstrated in this work as a way to reduce the hidden layer size of the neuron network. This algorithm could be more useful in questions with limitation number of neurons because it works effectively to prune some redundant neurons of a trained network while keeps the high accuracy and low loss, but it can hard to find how many neurons are necessary or find a minimum size of the hidden layer. We also used a fuzzy signature to analyse the dataset, which also performs well.

There are rooms for improvement as destituteness pruning only sets the pruned neurons' weight to 0 but does not remove them from the network indeed, which means, the size of the hidden layer does not change [5]. Therefore, time-consuming and calculate cost cannot be reduced significantly. Also, limited by the dataset, the advantage of fuzzy logic could not show clear, and the difference between methods is small. We could use a more complex dataset or reduce the dataset competence to challenge the algorithm.

5 References

[1] S. Kajan, D. Pernnecky and J. Goga, "APPLICATION OF NEURAL NETWORK IN MEDICAL DIAGNOSTICS".

[2] HK. Centre for Health Protection, "Severe Acute Respiratory Syndrome (SARS)," [Online]. Available: https://www.chp.gov.hk/en/healthtopics/content/24/47.html.

[3] T. D. Gedeon and D. Harris, "PROGRESSIVE IMAGE COMPRESSION," *International Joint Conference on Neural Networks,* no. 4, pp. 403-407, 1992.

[4] B. S. U. Mendis, T. D. Gedeon and L. Tóczy, "Investigation of Aggregation in Fuzzy Signatures," in *3rd International Conference on Computational Intelligence*, Singapore, 2005.

[5] D. Blalock, J. . J. Gonzalez Ortiz, J. Frankle and J. Guttag, "WHAT IS THE STATE OF NEURAL NETWORK PRUNING?.pdf," 6 May 2020. [Online]. Available: https://arxiv.org/pdf/2003.03033.pdf.

[6] K. W. Wong, T. Gedeon and L. Koczy, "Construction of Fuzzy Signature fiom Data: An Example of SARS Pre-clinical Diagnosis System," *Proceedings of IEEE International Conference on Fuzzy Systems,,* vol. 3, pp. 1649-1654, 2004.

[7] B. Mendis, T. Gedeon, J. Botzheim and L. Koczy, "Generalised Weighted Relevance Aggregation Operators for Hierarchical Fuzzy Signatures," in *International Conference on Computational Inteligence for Modelling Control and Automation and International Conference on Intelligent Agents Web Technologies and International Commerce (CIMCA'06)*, 2006.

[8] T. Vamos, L. Koczy and G. Biro, "Fuzzy Signature inn data mining," in *Proceedings Joint 9th IFSA World Congress and 20th NAFIPS International Conference*, 2001.