Feature Selection for Artificial Neural Network on Language Speaker Classification

Yu Yang

″Research School of Computer Science, Australian National University"″

u6412985@anu.edu.au

Abstract. Researchers use large datasets to ensure all the important data is fed to neural networks and the networks are larger. This paper discussed whether the size of networks. The evolutionary algorithm is used for feature selection and one reduction technology is used for reducing the size of hidden units. The results show that the feature selection can contribute to reduce the training epochs and reduction technology can reduce the size of hidden units. The parameters of the evolutionary algorithm are also discussed.

Keywords: Neural Network, Reading Distraction, Reduction Technology, Evolutionary Algorithm

1 Introduction

1.1 Background

Current approaches for classification use Artificial Neural Network (ANN) can achieve good results. The development of computation abilities of computers allows researchers to use datasets with more features and samples and the larger dataset can help make better results on the image classification [5]. There are more and more large dataset appears and the ANN is larger than before. Although, the ANN performance better than before, there are two questions left. One is whether all of the features collected in the dataset are useful. The other is whether all of the units in the ANN contribute to the final result.

The second question was discussed in the assignment one. The project discussed the influence of parameters of the reduction technology proposed by Gedeon & Harris in 1991. The results show that the reduction technology can significantly reduce the redundant units and keep the accuracy with proper parameters. That means, if the reduction technology applied, the memory cost and running time of the model will decrease. Then, if the size of the neural network is large, the reduction technology can save machine resources.

To reduce the size neural network, the size of the input also can be considered. The current dataset includes many features to ensure that all the important features are collected. Therefore, there are possibilities that some features are redundant. Jirapech-Umpai's and Aitken's research in 2005 show that evolutionary algorithms for feature selection can improve the stability of ANN performance [3]. This paper will apply the evolutionary algorithm to select features for the neural network.
The rest of this paper is organized as follows. Section two discusses the evolutionary algorithm for feature selection and the parameters of ANN. Section three discusses the influence of evolutionary algorithms and whether it is useful. The last section, section four, summaries the work of this paper and give future directions.

1.2 Dataset

The dataset is from the Australian National University collected by Copeland & Gedeon in 2015 [1]. There are 23 features in the dataset and 66 participants. The data relating to the humans' attention when reading easy and hard articles. The 23 features are listing below. The L1/ L2 is the class we want to classify, where L1 represents people whose mother tongue are English and L2 represents people whose mother tongue are other languages.

Feature	Data Type	Feature	Data Type
Participant ID	string	ratio of fixation duration in text are to out of text area	number
Condition	string	number of distractions Num fixations in text area	number
Text Type	string	Reading ratio Num fixations out of text area	number
Condition	string	skim ratio Num fixations in text area/out of text area	number
L1/L2	string	fixation duration in text area	number
Total num fixations	number	Longest reading sequence	number
skim ratio	number	Time Taken	number
number of distractions	number	Number of Distractions (from DB)	number
Reading ratio	number	Total Score	number
Total fixation dur (s)	number	Do you often use social media, email and/or instant message while you are reading course materials or work materials?	number
fixation duration out of text area	number	Do you find that you are distracted by these technologies during study or work time?	number
scan ratio	number		

Table 1. The dataset features

2. Methodology

We start with evolutionary algorithm to generate features (genes) for ANN. Then test the fitness values and replace the worse genes with offspring. After a center number of iterations, the best gene will be picked up and used for ANN for future training. The parameters of ANN are discussed to find a better parameter combination. To ensure the final performance is stable, cross validation is applied. Finally, the reduction technology is applied.

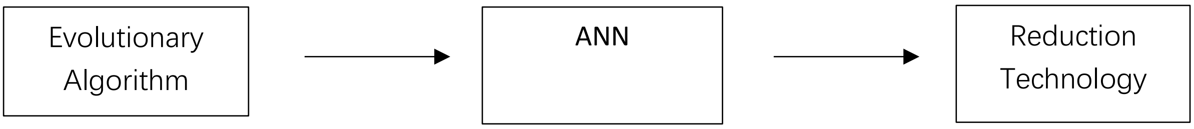

Fig 1. The whole process

2.1 Feature Selection

The evolutionary algorithm is an optimization technology based on stochastic search. It is inspired by evolutionary of creatures. Creatures inherit genes from their parents and good genes combinations have higher possibilities pass to the next generation. The gene has mutation possibilities and this can contribute to creating new gene combinations.

In this paper, feature combinations represent the gene combination. An array with 20 elements represents the feature combination, like figure 2. 0 means the feature is not selected and 1 means the feature is selected.

0	1	2	3	4	5	6
0	0	0	1	1	0	1

Fig 1. gene representation

We start with randomly initialise the genes and then replace the worse gene feature combinations with new offspring. The figure 3 descript the pseudo code.

1. Initialise the **N** gene combination randomly.
2. Generate new **OSN** offspring by crossover
3. Apply the mutation for each new offspring with a mutation rate.
4. Evaluate the fitness values of parents' gene combinations and replace the worse **OSN** parent with offspring.
5. Repeat the step 2 to step 5 until the satisfy the constrain.
6. Return the best gene combination

N is the total number of samples (gene combinations)
OSN is the number of offspring

Fig 2. Evolutionary Algorithm

When generating the offspring, the new gene combination randomly chooses the gene from its parents. Each gene in the new combination has a possibility to mutate to create new gene combinations. The fitness value is scored by an ANN. Each feature (gene) combination will be used for the same ANN training and then test the performance of the ANN. The accuracy value is regards as the fitness value. If the accuracy value is higher, the features contribute more in the final result. If the accuracy is lower, the features are useless for this problem. Thus, the better feature combination will be saved and worse will be dropped. The algorithm is convergence.

Some parameters of the evolutionary algorithm will be tested to evaluate their influence. The parameters include the mutation rate, the number of offspring and the number of iterations. On the other hand, some parameters are fixed. The length of gene combination is 20. The predictor ANN includes uncertain number of inputs depends on the inputs and 30 hidden units and 2 output units. The learning rate is 0.01 and the loss function is cross-entropy. The optimiser is stochastic gradient decrease. This ANN is a predictor to find a better gene combination. More discussion about parameters of ANN will be provided in the section 2.2.

2.2 Artificial Neural Network

2.2.1 Data Preprocessing

As table 1 shows, some data in the dataset is string, so it is important to convert those string data into numbers. First of all, the participant ID is an index which has no contribution to the classification and the first Condition is the combination of the Text Type and the second Condition. Thus, the participant ID and the first Condition columns can be dropped. The L1/ L2 column is the class we want to classify.

Next, normalization is applied. Some data are numbers, but the ranges are different. Some values are between 0-1 and some values are between 0-1000. Sola and Sevilla found that normalisation is crucial to obtain good result, especially for complex data, in 1997 [6]. In this paper, we normalise the continue numbers into 0-1 with the equation 1.

$$\mathrm{x} = \frac{x - min(X)}{max(X) - min(X)} \tag{1}$$

2.2.2 Neural Network Parameters

After the evolutionary algorithm returns the best feature combination, we use those features to train a better neural network according to the accuracy. The neural network includes three layers, inputs, hidden, outputs. The number of inputs is the number of the features. The number of hidden units is 30 larger than inputs and less than 2 times inputs [3]. There are two outputs units.

The learning rate and back propagation methods also influence the performance of the neural network (Zeiler, 2012) [7]. We test different learning rates. Then, we choose the best learning rate combination and train a new network. To improve the stability, the cross validation is introduced in this step.

2.3 Reduction Technology

Although, the best parameters are identified, it is not sure whether the hidden units are redundant. To reduce the size of the network, the reduction technology proposed by Gedeon & Harris in 1991 is introduced. The weights of the units can be regards as vectors and if the vectors are approximately parallel, those units can be merged or removed [2].

Condition	Action
The degree between vectors are less than 15	Merge those units, adding one weight to the other.
The degree between vectors are greater than 165	Remove both of them units

Others	Do nothing

Table 1. Network Reduction

3. Result and Discussion

3.1 Evolutionary Algorithm

We start with test the number of offspring in each iteration with other parameters fixed. The 25%, 50%, 75% parents will be replaced/ by the offspring in 3 tests, respectively. If there are less parents kept in each iteration, that mean more good gene combinations are dropped, because we cannot ensure that the next generations are better than those parents. The average precision is recorded in each iteration, Figure 4. The result shows that the accuracy increase slower with the lower number of offpring and higher number of offspring. When 50% number of parents are replaced, the average accuracy increases faster and the final result is better

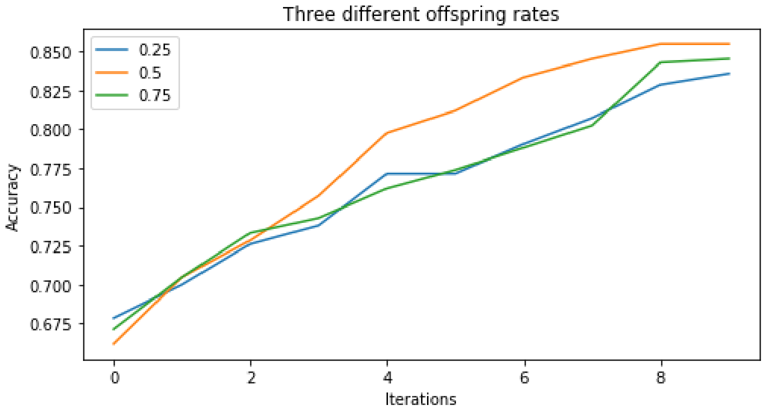

Fig 3. Test results with different offspring rates

Mutation rate is another factor that can influence the evolutionary algorithm. We compare the mutation rates among 0.01, 0.1, 05. If the mutation rate is higher, the gene combinations will not be stable and the good gene combinations are more likely to change. The Figure 5 shows the average accuracy in each iteration with different mutation rates. The precision fluctuating rise, but after ten iterations, the precisions are worse than the precisions in Figure 5. When the mutation rate is greater, the precision line is more unstable and the final result is worse.

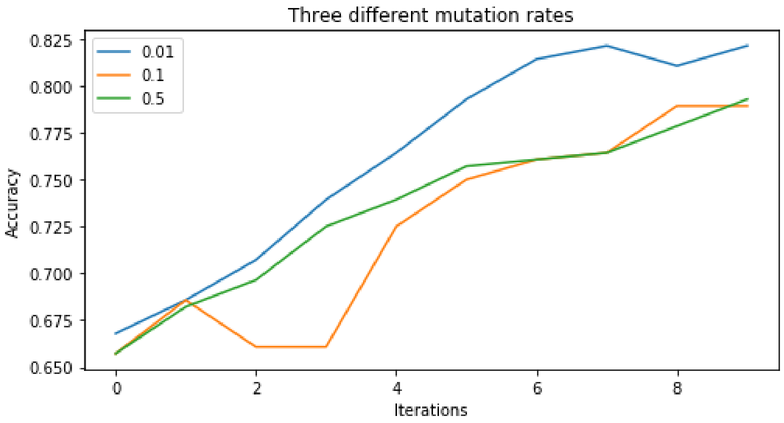

Fig 5. Test results with different mutation rates. Offspring rate is 0.5

We fix the offspring rate as 0.75 and the mutation rate as 0.01. Then, we select the features again.

3.2 Neural Network

The number of samples in the dataset is only 66. To ensure that the network result is reliable. We introduce the K-fold cross validation. We randomly train and test the network for K times which is 5 in this paper and we calculate the average of 5 accuracy scores. In addition, the learning rates we test are 0.01, 0.1 and 0.5. The results are listing below. When the learning rate is smaller, the loss decreases slower and when the learning rate is larger, the loss decreases faster. However, even the loss on training set is lower, the final accuracy is lower. One reason may be overfitting.

Accuracy\ learning rate	0.01	0.1	0.5
Fold 1	64.29%	57.14%	57.14%
Fold 2	69.23%	76.92%	69.23%
Fold 3	53.85%	76.92%	61.54%
Fold 4	76.92%	61.54%	53.85%
Fold 5	76.92%	61.54%	69.23%
Average	68.24%	66.81%	62.19%

Table 2 Cross Validation Accuracy

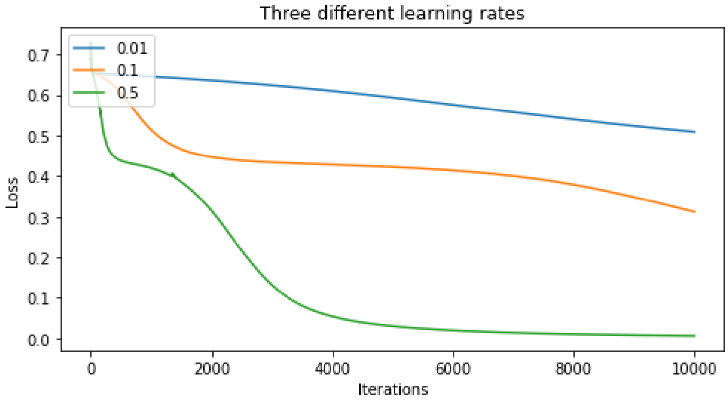

Fig. 4. Training loss with different learning rates

Therefore, the learning rate is fixed as 0.01 and a new train neural network is training. The accuracy is 69.23% on the test set after ten thousand epochs. The number of hidden units is 30. To test whether the size of hidden units can be reduced or not. The reduction technology is applied and 19 units out of 30 units are removed. The accuracy of new model is still 69.23% on the test set. When all the features are used for training and the number of epochs is the same. The accuracy is 64.29% on the test set and after reduction, the accuracy is 57.14%. The reason for the difference is the number of epochs is not large enough and the model is not convergent.

4. Conclusion and Future Work

The results show that feature selection can contribute to find good features for training and reduce the input size. In this way, the computation resources and time can be saved. This is useful when there are gigabytes data for training. The mutation rate will affect the stability for the gene combination, 0.01 – 0.1 is better. In this case, there will be enough new gene combination in each iteration. In addition, learning rates can affect the training speed, but larger learning rate can lead to overfitting. The range between 0.01- 0.1 good choices. Although the evolutionary algorithm reduces the sizes of the input, the reduction technology can also work the reducing the hidden units' size.

The future work can test the evolutionary algorithm on larger dataset or datasets in other areas. The dataset in this paper is small.

Reference

1. Copeland, L., & Gedeon, T. (2015, December). Visual Distractions Effects on Reading in Digital Environments: A Comparison of First and Second English Language Readers. In *Proceedings of the Annual Meeting of the Australian Special Interest Group for Computer Human Interaction* (pp. 506-516). ACM.

2. Gedeon, T. D., & Harris, D. (1991). Network reduction techniques. In *Proceedings International Conference on Neural Networks Methodologies and Applications* (Vol. 1, pp. 119-126).

3. Jirapech-Umpai, T. and Aitken, S., 2005. Feature selection and classification for microarray data analysis: Evolutionary methods for identifying predictive genes. *BMC bioinformatics*, *6*(1), p.148.

4. Karsoliya, S. (2012). Approximating number of hidden layer neurons in multiple hidden layer BPNN architecture. *International Journal of Engineering Trends and Technology*, *3*(6), 714-717.

5. Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. "Imagenet classification with deep convolutional neural networks." In Advances in neural information processing systems, pp. 1097-1105. 2012.

6. Sola, J. and Sevilla, J., 1997. Importance of input data normalization for the application of neural networks to complex industrial problems. IEEE Transactions on nuclear science, 44(3), pp.1464-1468.

7. Zeiler, M.D., 2012. ADADELTA: an adaptive learning rate method. arXiv preprint arXiv:1212.5701.