Interaction of Model Architecture, Model performance and Dropout in Constructive Cascade Network

Cheng Zhu

Research School of Computer Science
Australian National University
ACT 0200 Australia

U6456469@anu.edu.au

Abstract: Compared with the normal feedforward neural network backpropagation, the constructive cascade algorithm can automatically specify the size and topology of the network to be used, which can improve the efficiency and accuracy of model building by reducing the manual tuning process. This paper applies the constructive cascade algorithm to a multi-classification problem with three different architecture and different dropout size to evaluate the performance of the network. After analysis it shows the second cascade architecture that with Five neurons in a cascade layer with dropout equal to 0.3 get the best overall result which is 75.56%.

Keywords: Constructive cascade networks, neural network, feed-forward, backpropagation, dropout, model architecture

1. Introduction

The classification problem will always attract most of the attention in the field of neural networks. Multi-label classification is a recurring task in data science, and one of the current directions of neural network research is to solve multi-label classification problems more efficiently and accurately in the architecture of existing feed-forward neural networks. In this paper we use a set of assessment marks in an undergraduate Computer Science course [6], the goal is to predict the final grade of this course for each student.

Constructed cascade neural network is a feed-forward network that determine its own size and build the network structure for adaptive matching to a selected task [1]. In machine learning models, if the model has too many parameters and too few training samples, the trained model will easily produce the phenomenon of overfitting. Overfitting is often encountered when training neural networks. Overfitting is manifested in the following ways: the model has a small loss function on the training data and the prediction accuracy is high; however, the loss function on the test data is relatively large and the prediction accuracy is low. Dropout can be more effective in alleviating the occurrence of overfitting and achieving regularization to a certain extent.

The task of this paper is to propose constructive cascade algorithm with dropout to prove that a constructive cascade neural network could enhance the performance of a neural network when dealing with a multiple classification task. In this paper there are three different cascade network and we have applied different dropout parameter to in each experiment to avoid overfitting.

2. Method

2.1 Neural Network Topology

The algorithm used in this paper is slightly different from the constructive algorithm like cascade correlation (CasCor) or CasPer algorithm. It is same with the constructive algorithm in [3], that construct a basic network where all the input layer fully-connect to the output layer with no hidden layer and train with backpropagation algorithm [2].

Then add a number of cascade chunks (cascade layers) one by one, each of a fixed size set prior to training [3]. Every new cascade layer will fully-connect to the input layer and all pre-existing cascade layer and output layer and start training, and every new cascade layer input weight will be frozen after training finished (like in CasCor) and add new cascade layer into the network.

Different from the original CasCor, the training algorithm for cascade layers will not use the correlation between the new unit output and the residual error signal of the network to train cascade layers. The limitation of this correlation measure is that it will force the hidden neurons to saturate, which will eventually have a bad effect on the output of the neural network [4], the algorithm will use backpropagation instead [5].

There will be three different architecture of cascade network: the first one will have three neurons in each cascade layer, the second one will have five, and the third one will have ten. In this paper it will also provide the performance of original cascade network as comparison.

2.2 Data Set

The chosen data set is a set of assessment marks in an undergraduate Computer Science course [6], with the information about each student: the current program the student is enrolled in, which semester the student taken this course and the tutorial the student enrolled in, etc. The goal is to predict the final grade of this course for each student (Multiple classification tasks in supervised learning).

The assessment marks provided in the dataset combine to yield only 40% of the total mark. The remaining 60% is the final examination mark which is omitted. The aim of the task is to use the 40% weights of class assessment marks and students' information to predict their final grade of this course.

The size of the original data set is 152 with 15 features (where student ID has been dropped) and 1 output (final grade), 10 of the features (assessment marks) and the final grade are regular numeric features, other 5 of the features (student information) are categorical features.

The data has been shuffled and any sample that only provides basic student information without any marks has been dropped. Numeric missing features (assessment marks) have been replaced by 0, and there are no missing data in categorical features after dropping.

All numerical features have been normalized. The categorical feature that has more than 10 categories have been dropped in order to avoid overfitting, others will be converted into numeric data by one-hot encoding.

Final grade converts into the following output (as the output is ordinal data after converting, thus use label encoding instead of one-hot encoding):

- Fail, being a mark less than 50, represented by output 0.
- Pass, being a mark between 50 and 64, represented by output 1.
- Credit, being a mark between 65 and 74, represented by output 2.
- Distinction or above, being a mark of 75 or greater, represented by output 3.

2.3 Training Methodology

Training applied K-fold cross-validation (where K=10, data shuffled and randomly selected) to obtain reliable and stable models. The activation function used in all networks is rectified linear unit function (ReLu), the ReLu function enables more efficient gradient descent and back propagation, it avoids gradients explosion and gradient disappearance problems [7]. The learning algorithm used throughout the module is backpropagation (BP) algorithm. Loss function applied cross-entropy loss function, as this loss function performs better in multi-classification task [8].

Optimisation function is using an adaptive gradient, it can adjust different learning rates for each different parameter, updating frequently varying parameters in smaller steps and sparse parameters in larger steps. [9]. The module has added L2 regularisation [10] parameters which is set to 0.01 to avoid overfitting.

The network will freeze the current training cascade layer and add a new cascade layer when the decreased value of loss (loss value in previous epoch minus loss value in current epoch) is less than 0.0001 and larger than 0. The learning rate is set to 0.0025. The number of epochs is set to 1000. The module will take the test accuracy as the result to evaluate the performance of the module.

The network has also applied dropout as an important hyper-parameter. The details of dropout setting will be described in Evaluation session, basically it will be set as 0.3, 0.5, 0.7, respectively.

3.Evaluation and Discussion

We have evaluated our method using the data set as described in Section 2.2. The problem to be solved is to predict final marks with limited information and assessment marks.

We will use the test accuracy to compare the performance of the cascade network with a baseline neural network and the best result performed in [6].

The baseline is a network with one hidden layer, feed-forward with learning rate set to 0.0025, number of input neurons is set to 20, the number of epochs is 500 and L2 regularisation parameter is setting to 0.0545, all the other settings is the same as described in section 2.3, except the part of cascade settings.

There are 3 different constructive cascade networks with cascade layer sizes set to 1, 3, and 5 separately, and the network is an original cascade network when cascade layer size is equal to 1.

All the result is using the average performance of the 20 repetitions of each experiment,

Table 1 shows shows average training and testing loss for each experiment, and the total number of hidden (cascade) neurons (thus the average number of cascade layers equal to the total number of hidden neurons divided by corresponding cascade layer size).

Table 1. Average of performance on data set without dropout

Network	Train Accuracy	**Test Accuracy**	Train Loss	Test Loss	Total number of hidden (cascade) neurons
Orign Cas (Cascade layer size =1)	77.91%	65.94%	0.6210	0.7701	11
Cas 3 (Cascade layer size =3)	84.56%	**72.52%**	0.4334	0.7010	14
Cas 5 (Cascade layer size =5)	86.55%	70.89%	0.3884	0.6512	29
Cas 10 (Cascade layer size =10)	87.14%	70.14%	0.3744	0.7112	47
Baseline	69.45%	65.55%	0.8251	0.8742	20
Network in [6]	75.0%	66.0%	None	None	5

As shown in table 1, Baseline and network in [6] have a fairly close test accuracy (66.55% and 66%), and the constructive cascade network has a relatively better performance overall compared with baseline network and network in [6], together with the cascade network has a relatively lower loss on both training and testing.

The original cascade network Origin Cas does not perform any better compare with Baseline and Network in [6], and Cas 3 get the best result that has a 6.97% higher test accuracy than the baseline network, with a

Table 2. Average of performance on data set with dropout

Network	**Test Accuracy (Dropout=0.3)**	Test Accuracy (Dropout=0.5)	Test Accuracy (Dropout=0.7)	Average number of hidden (cascade) neurons
Cas 3 (Cascade layer size =3)	73.34%	71.16%	56.45%	22
Cas 5 (Cascade layer size =5)	**75.56%**	72.24%	54.5%	31
Cas 10 (Cascade layer size =10)	70.1%	63.34%	61.33%	75

significantly lower loss on training and testing, and the total number of hidden neurons is also lower. A larger cascade layer (Cas 5,10) also results in a lower test accuracy compare with Cas 3.

Base on the performance in Cas 3, Cas 5, and Cas10, a constructive cascade network could have a better performance compared with a regular network, but with a larger difference between training and test accuracy, it indicates that constructive cascade has increased the accuracy of the result but also shows a relative lack of generalization performance and would be easy getting into a situation of overfitting. When the size of the cascade layer increasing (Cas 3,5,10), the tendency of overfitting is accentuated.

The total number of hidden neurons also proves the existence of overfitting from the side, the number of hidden neurons in Cas 10 have 47 average hidden neurons, it is inevitable that so many hidden neurons (as opposed to baseline) will fall into overfitting. But one of the advantages of the cascade network is the network determines its own size and topologies, which will reduce the chance of overfitting [11].

Table 2 shows shows the average of performance on data set with dropout, from the table it shows that when dropout equal to 0.3, Cas 5 get the best overall performance, which is better than the result in Cas3, table 1. The table here shows that when adding dropout parameter, the average number of hidden cascade neurons has increased by 5-10, and with the increasing number of dropout size, the overall accuracy is also decreasing.

From the table above it shows that when add dropout parameter, the performance of the constructive cascade network is better than without dropout settings, and Cas 5 get the best result, which is also different from the result in table 1, it shows that dropout indeed increase the average performance of the constructive cascade network, but with increased complexity (the increasing number of hidden neurons).

4. Conclusion and Future Works

This report has introduced the algorithm for constructing cascade networks with dropout for multiple classification task using cascade layers, by constraining the cascade process by adding fixed size cascade layers. In evaluation, it shows the second cascade architecture that with Five neurons in a cascade layer with dropout equal to 0.3 get the best overall result which is 75.56%.

Future work will focus on the automatic determination of the number of hidden neurons in each cascade layer under the architecture of constructive cascade network to enhance generalization performance.

References

1. wok, T.-Y., Yeung, D.-Y.: Constructive Algorithms for Structure Learning in Feedforward Neural Networks for Regression Problems. IEEE Trans. on Neural Networks 8, 630– 645 (1997)
2. Fahlman, S. E., & Lebiere, C.: The cascade-correlation learning architecture. CARNEGIE-MELLON UNIV PITTSBURGH PA SCHOOL OF COMPUTER SCIENCE (1990).
3. Khoo, S., & Gedeon, T.: Generalisation Performance vs. Architecture Variations in Constructive Cascade Networks. In International Conference on Neural Information Processing, pp. 236—243. Springer, Berlin, Heidelberg (2008).

4. Treadgold, N. K., & Gedeon, T. D.: A cascade network algorithm employing progressive RPROP. In International Work-Conference on Artificial Neural Networks, pp. 733--742. Springer, Berlin, Heidelberg (1997)
5. Goyal, S., & Goyal, G. K.: Cascade and feedforward backpropagation artificial neural networks models for prediction of sensory quality of instant coffee flavoured sterilized drink. Canadian Journal on Artificial Intelligence, Machine Learning and Pattern Recognition, 2(6), 78--82 (2011)
6. Choi, E. C. Y., & Gedeon, T. D.: Comparison of extracted rules from multiple networks. In Proceedings of ICNN'95-International Conference on Neural Networks,Vol. 4, pp. 1812--1815. IEEE (1995)
7. Xu, B., Wang, N., Chen, T., & Li, M.: Empirical evaluation of rectified activations in convolutional network. arXiv preprint arXiv:1505.00853 (2015)
8. Nam, J., Kim, J., Mencía, E. L., Gurevych, I., & Fürnkranz, J.: Large-scale multi-label text classification—revisiting neural networks. In Joint european conference on machine learning and knowledge discovery in databases, pp. 437—452, Springer, Berlin, Heidelberg.(2014)
9. Kingma, D. P., & Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
10. Bilgic, B., Chatnuntawech, I., Fan, A. P., Setsompop, K., Cauley, S. F., Wald, L. L., & Adalsteinsson, E.: Fast image reconstruction with L2 ‐ regularization. Journal of magnetic resonance imaging, 40(1), 181--191(2014)
11. Tetko, I. V., & Villa, A. E.: An enhancement of generalization ability in cascade correlation algorithm by avoidance of overfitting/overtraining problem. Neural Processing Letters, 6(1), 43-50 (1997)