Multi-class Classification: Constructive Cascade and Long Short-Term Memory Neural Networks with Genetic Algorithm for Feature Selection

Yuchong Yao

Research School of Computer Science, Australian National University ACT, Australia
U6307906@anu.edu.au

Abstract. Classification problems are one of the most common research domains in machine learning and have great impact on people's daily life. Neural Networks [1] have been proved to be powerful tools for solving classification problems and are still active in all kinds of research areas. However, Neural Network classifiers depend on proper structures and topologies in order to achieve satisfiable results, which is still a difficult problem. Improper sized network can cause underfitting or overfitting, which damages the generalisation and other performance. Besides, traditional feed forward Neural Networks show incapability of performing classification on time series data [4] as they have no notion of order in time. In this paper, Neural Network techniques have been applied to predict participants' vote about whether an image has been manipulated based on a set of time series eye gaze data [19]. Constructive Cascade Algorithm (CCA) [2] has been introduced due to its ability to decide the network structure during training, which results in better generalisation [16][20]. Long Short-Term Memory (LSTM) [6] has been applied as it overcomes lags of unknown duration between events in time series and traditional gradient problems for other Recurrent Neural Networks. Genetic Algorithm (GA) [10] has been used for feature selection to determine the most relevant subset features for classification [13]. The results confirm that Constructive Cascade has good generalisation ability and gives better network structure for the given problems, which can be applied later for other techniques. LSTM performs better on time series data in terms of losses, accuracies and learning. GA provides feature subset for the classifier, which removes less relevant features and results in better pattern recognition and learning effectiveness. Those techniques have shown powerful performance on classification problems.

Keywords: Classification, Neural Network, Feed forward, Constructive Cascade Neural Network, Recurrent Neural Network, Long Short-Term Memory, Genetic Algorithm, Evolution, Feature Selection

1 Introduction

Classification problems on time series data have become one of the most active fields in machine learning research areas. Various models and approaches have been developed in response to these problems. The problem described in this paper is using eye gaze data collected during the experiments to determine participants' ideas about whether the subject image has been manipulated [19], which is a typical instance of classification problems.

Traditional feed forward networks with back-propagation have shown reasonable performance in various kinds of tasks, but often fails due to the limitation of determining the optimal network structure. Undersized or oversized models results in problems like underfitting or overfitting, which damages the performance for classification. The problem motivates the presence of Constructive algorithms [2], whose network is usually initialized with minimal structure and incrementally build on layers and hidden units during training to discover satisfiable structure [17]. Constructive Cascade Algorithm has been applied in this paper which aims to discover the preferred network topologies and generalisation property.

Due to problems such as lags of unknown duration between events, feed forward networks often fail to give desirable results for time series prediction. Therefore, it motivates the application of recurrent neural networks (RNN), which are distinguished from feed forward networks for their concept of memory as they not just use the current input, but also what they have perceived previously in time. Long Short-Term Memory is introduced in this paper, which aims to overcomes the vanishing gradient problems of RNN by employing multiplicative gates that enforce constant error flow through the internal states of special units called memory cells [7], while also preserves the power of RNN on processing time series data.

As the size of data is exponentially growing, training models can be computationally expensive and cost a lot of time. Besides, irrelevant features in the dataset prevent the model from recognising intended patterns for classification. Therefore, Genetic Algorithm is adopted in this paper for feature selection, which will choose the best subset of features in the initialized population through crossover, mutation and selection [12] and then applies those selected features to build the model. In this paper, GA aims to help generate models to find most relevant patterns and make training more effective in classification problems.

2 Method

2.1 Data Set

There are two sets of data used in this paper, which are proposed by Caldwell et al. [19] on her experiments about eye gaze tracking for participants looking at a set of unmanipulated and manipulated images. The first set of data contains 372 entries and each entry has 8 features. Features include participant ID, set of information about eye gaze, image information and participant vote. The second set of data has 31114 data entries with 9 features. It is a time series data containing more detailed information about fixation, fixation position, image, time and samples. The dataset used for building models is the combination of these two data, which merges them according to participant id and image id. There are 9 features in this combined dataset coming from the second data and additional vote information from the first data.

The classification task aims to recognise patterns from the given features and determine the participants' vote for whether the image is manipulated or not. Therefore, subset of 9 features are treated as inputs and the vote is the model output.

For inputs, they are of different ranges and the value of the input can affect weights during the learning process. Data normalisation methods have been applied for the purpose of better modelling and faster learning [3]. The normalisations adopted are Student's t-statistic (1) and Min-Max Feature scaling (2).

$$x = (x - x.min) / (x.max - x.min) \quad (\mathbf{1})$$

$$x = (x - x.mean) / x.std \quad (\mathbf{2})$$

The data is divided into training set, validation set, and test set according to the image ID. Hence, data of one image will be used for validation, data of another image will be used for testing and the rest of data for other images will be used for training. Data is sorted based on start time and participant ID in order to model the sequence for LSTM.

2.2. Evaluation Method

In this paper, loss and accuracy are used to describe the performance of the generated model during training, validation and testing. Cross-Entropy (CE) loss is adopted in the experiment as it yields

better results in terms of probability distribution and multi-class classification. Accuracy is calculated by the portion of correct classified patterns over all patterns.

2.3 Constructive Cascade

The network is initialised with minimal structure and hidden units are added incrementally to the network until a satisfactory solution is found. Similar to Layered Casper [21], a cascade layer containing fixed number of neurons is added each time and connects to all inputs and the previous hidden layers [18] [22] as shown in Figure 1. This can speed up the convergence and learning process while also helps to escape from local minima as new layers increase the dimensionality of error surface, which allows the network to continue reducing error [16]. According to paper proposed by Khoo and Gedeon [18], hyperbolic tangent and resilient back propagation (RPROP) are applied to the model for faster convergence and can use only the sign of gradient to adapt step size which leads to smaller network and better generalisation [15].

Figure 1 Topology

In this paper, network is initialized with only input and output layers with random weights and bias. Number of neurons each layer is set as 5 with 0.01 learning rate and 500 training epochs. A threshold for adding new cascade layer is set to 0.3 compared to the current loss of the model. The model will also attempt to add new layers when all 500 epochs are finished. A tolerance of 0.05 on validation loss between two consecutive models is applied to determine whether a termination condition is met and ensure the optimised structure.

2.4 Long Short-Term Memory

Long Short-Term Memory RNN is a special kind of RNN which is capable of learning long-term dependencies and solve the gradient vanishing problem for tradition RNNs [6]. LSTMs have chain like structures as RNNs, but the repeating modules have a different structure. The non-linear units in the hidden layer are replaced by memory blocks. Each memory block contains one or more memory cells along with input, output and forget gates which control flow of information into and out of the memory cell [5] as shown in Figure 2. It helps preserve the error that can be backpropagated through time and layers. By maintaining more constant error, it enables the network to continue to learn over more time steps and result in more successful runs and faster learning [6][8].

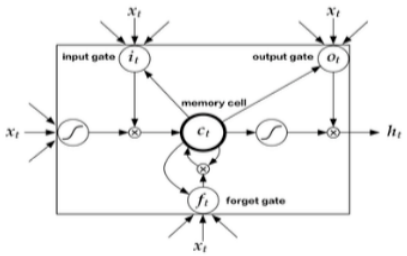

Figure 2 LSTM Memory Block

The LSTM model used in this paper is a stacked LSTM which has two recurrent layers [9], with the second LSTM taking in outputs of the first LSTM and computing the final results. Input dimension is set according to selected features by GA. Hidden states and cell states are initialized according to number of layers, batch and hidden dimension which also ensure the model to be stateful [9]. A linear output layer connects with LSTM layer to calculate final predict value. Hidden dimension is set to 5 in correspond to that of Constructive Cascade. The model adopts Adam optimiser with learning rate of 0.01 and runs for 500 epochs for better learning results in terms of speed and accuracy. The output layer has dimension of 3 which predict the participant vote for the given image.

2.5 Genetic Algorithm

This paper adopts Genetic Algorithm for feature selection in order to choose the best feature subset from the given data. By applying GA with the above techniques, it helps to improve efficiency and accuracy of the model and is able to remove the potential outliers as well [11][14]. A chromosome

(DNA) represents a feature subset (as shown in Figure 3), where 1 indicates the feature in the subset and 0 means the feature is not in the subset. Features in a subset are used as classifier inputs. The size of DNA is set to be 9, which is the number of all features in the data.

Figure 3 Chromosome

Population size is set to be 20 plus one DNA with all the features selected. Number of generations is set to be 20 for the GA. There is a trade-off in these settings as undersized population and generations will not ensure optimal solution, but oversized ones could be very computational expensive. Crossover rate and mutation rate is set to be 0.8 and 0.02 respectively to ensure the optimality and diversity of GA. Fitness of individuals is calculated based on model classification performance and individual will be selected based on its fitness [10] [12].

The crossover is to choose two parents within the selected pool with good fitness value to generate offspring. The rationale for this setting is to ensure the good parts of DNA can be preserved and fused to the next generation for optimisation. The replacement strategy used in this paper is similar to Parent-offspring method, where the better offspring will replace its parent in the population [11][12].

The stopping criteria is either there exists a DNA which meets the predefined requirements for fitness (e.g. 0.3), or all generations finished, then GA gives the best individual in the population [10].

3 Results and Discussion

The experiment in this paper will focus on three aspects. Firstly, experiments are conducted to show the generalisation property of Constructive Cascade network and observe its ability to deal with local minima problem. Secondly, the results will show benchmark comparison for LSTM and Constructive Cascade on time series data and investigate the advantages of LSTM over feed forward neural networks. Finally, there will be results proving the extra power offered by Genetic Algorithms on feature selection in terms of better modelling and less computational cost. In the paper proposed by Sabrina [19], there is 56% mean accuracy for participants identifying these images with correct labels (61.3% for unmanipulated and 50.1% for manipulated).

The results in Table.1 are the average of 5 best runs over 20 runs in total for 3 methods used in this paper.

Table 1. Overall Performance

Method	No. Layers	Test Accuracy	Test Loss	Train Accuracy	Train Loss
Constructive Cascade	3	31.62%	6.04	83.26%	0.31
LSTM	3	52.28%	3.01	85.07%	0.37
LSTM + GA	3	55.37%	1.83	84.26%	0.35

As shown in the table, the Constructive Cascade network performs about the same compared to the other two methods during training, however, when test set has been applied to it, Cascade algorithm gives very low test-accuracy (as bad as random guess) and high test-loss. While LSTM and hybrid (LSTM + GA) yield better results in both training and testing, which prove the power of LSTM on time series data and GA for better and more effective modelling.

3.1 Analysis on Constructive Cascade (CCA) and LSTM

During the experiments, CCA and LSTM show different properties on performing time series classification task. The CCA is very unstable during the test runs as the test accuracy can vary in a

very large range, while the LSTM is quite stable compared to CCA and generates better results. As shown in Table.1, there is no much difference between both models during training, however, LSTM did much better during on testing data, which proves the idea that LSTM learns the time series better in terms of long-time dependencies and shows the incapability of feed forward neural networks on modelling large scale of time series data. As shown in Figure 4, the green line indicates the loss for LSTM and the blue one is for CCA. It can be seen that LSTM learns steadily and reduces the loss to a relevant low level. While CCA learns very fast, but when a new cascade is added, loss becomes even higher than the loss during the initialisation. It indicates that the model did not learn the data well as it can only process with the current input and can't incorporate with previous memory.

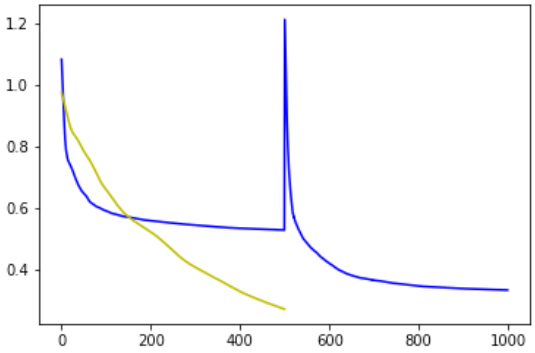

Figure 4 Training loss of CCA and LSTM

However, it is worth noticing that after the current model converges, the CCA continue reducing the loss by adding a new layer, which clearly proves its ability to escape from local minima. Also, according to my previous work [22] on comparing CCA with traditional multi-layer neural networks, CCA uses a smaller number of weights and keeps smaller network size compared to multi-layer networks of fixed topologies while CCA still has better results on classification. It further proves the generalisation property of CCA.

3.3 Analysis on LSTM and Hybrid method (LSTM + GA)

The experiment results show that there is no much difference between the performance of LSTM and hybrid methods in terms of loss and accuracy. In general, LSTM did slightly better during training, while hybrid method did better during the testing. As shown in Figure 6 and Figure 7, LSTM (red line) learns steadily and achieved lower loss and higher accuracy. While, the Hybrid method (green line) converge faster than LSTM but did not reduce the loss as low as LSTM.

It shows that Genetic Algorithms helps to achieve faster learning and improve the learning efficiency. It helps the model to take best subset of features as inputs, which results in better

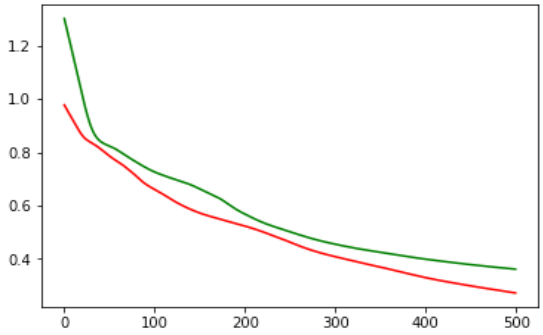

Figure 5 Loss for LSTM and Hybrid in training

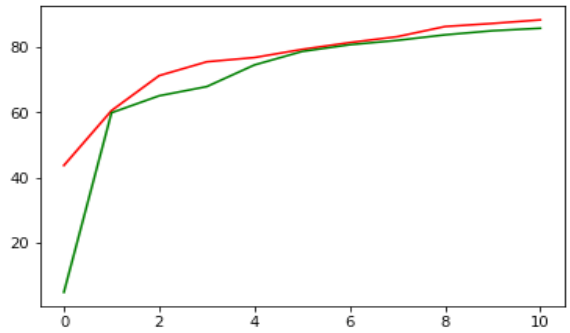

Figure 7 Accuracy for LSTM and Hybrid during training

generalisation for the given problems and better performance on pattern recognition.

In my experiments, the GA gave the best subset, which only contains 7 features compared to 9 features in the dataset. It offers great power in removing redundant or less relevant features, for example, duration is related to start time and end time, therefore, the duration feature can be possibly removed from the features set.

By applying less features for modelling, there are several advantages. Firstly, less training features means less computational power and time required for generating the model. Secondly, by choosing better subset of the features, model can learn better about the hidden nature of the dataset and recognise correct patterns, which lead to better test performance. Finally, GA can potentially remove

the outliers in the given data, as the select operator in GA keeps selecting better population and replace the individuals of bad performance. This enables the model to deal with more noisy inputs. However, the downside could be the extra computational power needed for performing feature selection using GA. Large population and generation require large amount of time and computational power to select the best individual. As the fitness of individual is coming from the classification performance using the selected features, there will be a number of models being generated in each generation using large amount of data. It will be lucky if best individual appears from crossover and mutation in the early generation, which meets the stopping criteria. However, sometimes GA has to go through all generations in order to generate the best chromosome. Therefore, there is a trade-off between the extra computational power needed for performing GA and the computational power saved when using best feature subset to model the data.

4 Conclusion and Future Work

This paper explores approaches for multi-class classification problem using Constructive Cascade Algorithm, Long Short-Term Memory and Genetic Algorithm. The Constructive Cascade network produces better structure and results in better generalisation compared to traditional multi-layer networks [22] and it also helps the model to escape from local minima. It can be used to explore preferred structure of network for the given problems and served as useful information for more advanced techniques. However, the downside of it is that Constructive Cascade Algorithms are computational expensive and not suitable for time series data. LSTM shows better results on modelling time series data compared to feed forward neural networks in terms of generation performance and learning efficiency. It is proved that recurrent neural networks and memory blocks offer LSTM with extra power to deal with long-time dependencies and other gradient problems. Genetic Algorithm has been shown to be very effective for feature selection and outlier detection. GA helps to choose best feature subset which not only helps the model to learn the data better, but also accelerate the training process as redundant and less irrelevant data are removed.

In the future, I will research on more effective initialization methods for Constructive Cascade as random initialization for weights and bias tends to be not stable enough for the given classification problem, hence, more problem-dependent strategies can be developed. Besides, LSTM still suffers from problems such as gradient clipping and reversing inputs and is not very stable. To solve the problem, further investigation can be focused on Attention layer [23], or research about ResNet [25] as LSTM shares some similarity with it (multiple switch gates bypass units to remember longer time step). Genetic convolutional strategy can also be taken into consideration as some results [24] indicate that a simple convolutional architecture outperforms canonical recurrent networks such as LSTMs across a diverse range of tasks and datasets. More effective and less computational expensive strategies can be explored for GA in terms of using less population, generation and datasets [14]. Different crossover (e.g. UNDX, SBX, Diagonal), mutation (e.g. random, in-order) and replacement strategies such as Tournament, Elitist or Conservative can be further investigated [26] and more complex combination of hybrid techniques could be implemented in the future.

References

1. Bishop C.M (1995) Neural networks for pattern recognition Oxford Univ. Press, Oxford 1995.
2. Parekh, R., Yang, J., Honavar, V.: Constructive neural-network learning algorithms for pattern classification. In: IEEE Transactions on Neural Networks, vol. 11, no. 2, pp. 436-451, March 2000.
3. Howley T., Madden M.G., O'Connell ML., Ryder A.G. (2006) The Effect of Principal Component Analysis on Machine Learning Accuracy with High Dimensional Spectral Data. In: Macintosh A., Ellis R., Allen T. (eds) Applications and Innovations in Intelligent Systems XIII. SGAI 2005. Springer, London
4. J.T. Connor, R.D. Martin, L.E. Atlas (1994) Recurrent neural networks and robust time series prediction IEEE Transactions on Neural Networks (Volume: 5, Issue: 2 , Mar 1994)
5. Has¸im Sak, Andrew Senior, Franc¸oise Beaufays (2014) Long short-term memory recurrent neural network architectures for large scale acoustic modelling 15th Annual Conference of the International Speech Communication Association Singapore September 14-18, 2014
6. Hochreiter S, Schmidhuber J (1997): Long short-term memory. Neural Computation Volume 9 | Issue 8 | November 15, 1997 p.1735-1780
7. Pankaj Malhotra1, Lovekesh Vig2, Gautam Shroff1, Puneet Agarwal1 (2015) Long short term memory networks for anomaly detection in time series ESANN 2015 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 22-24 April 2015
8. Bram Bakker (2002) Reinforcement learning with long short-term memory Proceeding NIPS'01 Proceedings of the 14th International Conference on Neural Information Processing Systems: Natural and Synthetic Pages 1475-1482
9. Correâ D.C., Levada A.L.M., Saito J.H.: Improving the Learning Speed in 2-Layered LSTM Network by Estimating the Configuration of Hidden Units and Optimizing Weights Initialization. In: Kůrková V., Neruda R., Koutník J. (eds) Artificial Neural Networks - ICANN 2008.
10. Yang J, Honavar V (1997) Feature subset selection using a genetic algorithm Iowa state university digital repository 5-3-1997
11. Leardi R (1994) Application of a genetic algorithm to feature selection under full validation conditions and to outlier detection Journal of Chemometrics vol. 8, 65-79 (1994)
12. W.F.Punch, E.D.Goodman, Min Pei, Lai Chia Shun, P.Hovland, R. Enbody (1993) Further Research on feature selection and classification using genetic algorithms Appeared in ICDA93, pg 557-564, Champain Ill
13. Vafaie H, Kenneth De Jong (1992) Genetic algorithms as a tool for feature selection in machine learning Proc. Of 1992 IEEE Int. Conf, on Tools with AI Arlington, VA, Nov 1992
14. Jain A, Zongker D (1997) Feature selection: evaluation, application, and small sample performance IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 19, No.2, February 1997
15. Treadgold N.K., Gedeon T.D. (1997) A cascade network algorithm employing Progressive RPROP. In: Mira J., Moreno-Díaz R., Cabestany J. (eds) Biological and Artificial Computation: From Neuroscience to Technology. IWANN 1997
16. Treadgold N, Gedeon T (1999) Exploring constructive cascade networks. IEEE Transactions on Neural Networks 10:1335-1350.
17. Tin-Yau Kwok, Dit-Yan Yeung (1997) Constructive algorithms for structure learning in feedforward neural networks for regression problems on IEEE Transactions on Neural Networks (Volume: 8, Issue: 3, May 1997)
18. Khoo S., Gedeon T. (2009) Generalisation Performance vs. Architecture Variations in Constructive Cascade Networks. In: Köppen M., Kasabov N., Coghill G. (eds) Advances in Neuro-Information Processing. ICONIP
19. Sabrina Caldwell, Tamás Gedeon, Richard Jones, Leana Copeland (2015) Imperfect understandings: a grounded theory and eye gaze investigation of human perceptions of manipulated and unmanipulated digital images Proceedings of the World Congress on Electrical Engineering and Computer Systems and Science (EECSS 2015)
20. W. Fang, R.C. Lacher (1994) Network complexity and learning efficiency of constructive learning algorithms on Proceedings of 1994 IEEE International Conference on Neural Networks (ICNN'94)
21. Tengfei Shen, Dingyuan Zhu (2012) Layered_CasPer: Layered cascade artificial neural networks on the 2012 International Joint Conference on Neural Networks (IJCNN)
22. Yuchong Y (2019) Constructive cascade for training feedforward neural networks ANU Annual Bio-Inspired Computing conference 2019
23. Xu K, Ba J, Kiros R, Cho K, Courville A, Salakhutdinov R, Zemel R, Bengio Y (2015) Show, attend and tell: neural image caption generation with visual attention eprint arXiv:1502.03044
24. Bai S, Kolter Z, Koltun V (2018) An empirical evaluation of generic convolutional and recurrent networks for sequence modelling eprint arXiv:1803.01271
25. He K, Zhang X, Ren S, Sun J (2015) Deep residual learning for image recognition The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770-778
26. Engelbrecht A.P. (2018) Computational intelligence: an introduction Hoboken: John Wiley & Sons, Ltd., 2008.