Casper Neural Network for Depression Level Classification under Genetic Algorithm based Feature Selection[⋆]

Zhiyong Sun

Australia National University
u6123044@anu.edu.au

Abstract. The level of depression can be determined by observing physical signs such as body temperature, pupils, or skin temperature. In this paper, we will use a genetic algorithm based feature selection method and investigate the Casper algorithm compared with the traditional fully connected neural network in diagnosing depression levels in terms of both accuracy and training speed. The results show that Casper is much inferior to fully connected neural network in terms of training speed, while the accuracy is similar. However, both algorithms showed around 80% accuracy in the initial determination of whether the depression was severe or not.

Keywords: casper· genetic algorithm · depression.

1 Introduction

Depression is the number one psychological factor that causes patients to commit suicide[1], timely diagnosis of depression can save many lives. However, detecting depression is a very complex and subjective task. Not only do untrained laypersons have a low rate of correct depression level ratings[2], but some professional scales have a high rate of misdiagnosis. In one study[3], the misdiagnosis rate using the Mini International Neuropsychiatric Interview (MINI) reached 65.9% for major depressive disorder, 92.7% for bipolar disorder, 85.8% for panic disorder, 71.0% for generalized anxiety disorder, and 97.8% for social anxiety disorder. Therefore, it is urgent to find a quick and effective way to diagnose depression.Neural network diagnosis has shown good performance when diagnose traditional diseases such as heart disease, In some studies the accuracy could be 90% or more[4]. Neural networks have two advantages over traditional diagnose through questioning or scale. The first advantage is speed. A neural network's classification of a piece of data can be seen as instantaneous, and ideally, the doctor will have the reference information from the algorithm before the traditional consultation begins. The second advantage is that there is no conflict between collecting physical data and doing emotional questions. This allows the neural network algorithm diagnosis to be used as an effective auxiliary diagnostic reference.

In this paper, we investigate whether Casper neural network has advantages in diagnosing depression level compared with ordinary feed-forward neural network, and because some features of the dataset are different statistical results from the same source, how to choose the appropriate subset of features is also the main problem of this paper. The maximal information coefficient (MIC) algorithm and a genetic algorithm based on the MIC algorithm are used for feature selection. Because the MIC processed data outperforms the unprocessed features in both Casper and feed-forward algorithms (the performance results will be shown in the experimental section), and we use its results as a baseline for comparison with the genetic algorithm afterwards. It will also be used for the generation of the chromosome of the genetic algorithm, which will be discussed in the methodology section.

1.1 Dataset

The main dataset used in this paper is Pupillary Dilation collected bt [2]. The dataset contains 4 target classes: 0 for no or slight depression, 1 for middle-level depression, 2 for strong level depression and 3 for grievous level depression. It contains 39 features, they are the statistical results of the original data including minimum, maximum, Mean, standard deviation, variance, root mean square and some other results. These data are a mix of 'filtered' and 'unfiltered' features, so some features are not normalized.

This dataset contains 192 pieces of data, 48 items for each category, which is a very balanced dataset. Therefore, there is no need to perform virtual data generation. However, since it is not the original data and has been used twice, some attributes are very strongly correlated with each other. Therefore, for this dataset, the goal of data preprocessing is to remove redundant

[⋆] Supported by Australia National University.

attributes to reduce unnecessary computations. This will be achieved by both MIC algorithm and genetic algorithm. Another feature of this dataset is that it comes from 13 people, each providing 16 pieces of data. To make full use of this information it is necessary to use the leave-one-out method [2]. This will be described in detail during the experimental phase.

2 Methodology

2.1 MIC algorithm

I used the maximal information coefficient (MIC) algorithm [5] to visualize the correlation between different attributes.

$$MIC(x, y) = \frac{I_{MIC}\{x : y\}}{Z_{MIC}}$$

I_{MIC} is calculated by the number of bins imposed on the x and y axes n_x, n_y, and $Z_{MIC} = log_2(min(n_x, n_y))$. The heatmap of original features is shown in figure 1:

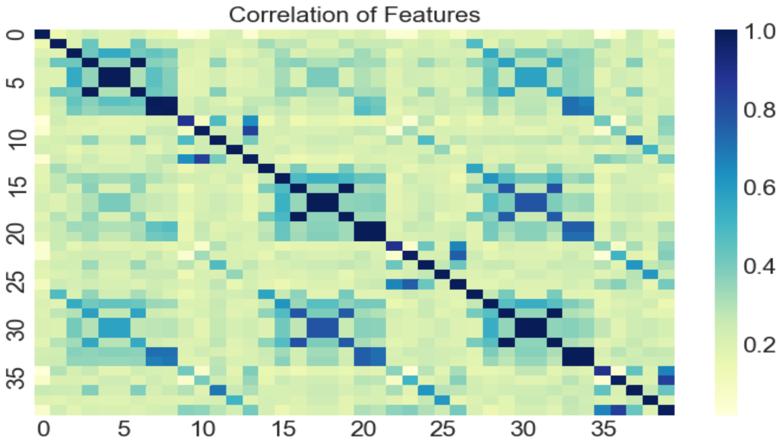

Fig. 1. The heatmap of correlation for original dataset

Ideally all the features should have low correlation with each other, all the grids except the diagonal should be green. The index of those feature groups that have a correlation close to 1 with each other is as follows: (3,4,5,6), (7,8,9), (16,19), (17,18), (19,20), (29,32), (30,31), (33,34). I kept only one feature in each group with correlation close to 1 and deleted the others. This results in a significant increase in training speed and almost no impact on accuracy. The performance before and after feature selection will be shown in the following experiment part. The dimension of the dataset after this feature selection is 192x25.

2.2 Genetic algorithm

Genetic algorithm is an algorithm that mimics natural selection process by gradually selecting the dominant individual among a random population of individuals, and then recombining the dominant individuals to produce a new population, thus searching for the desired result[6]. Individuals in the feature selection problem are binary arrays of equal length as maximum number of features, where 1 means the feature is selected and 0 means it is not selected. Thus each individual can represent a subset of features. There is no fixed answer for how to measure each individual's fitness. Since the number of individuals and iterations of genetic algorithms is relatively large, the algorithm for measuring fitness needs to be efficient. The KNN algorithm has been used for the computation of fitness in the study of geological information[7]. There are also studies that use the results of multiple conventional feature selection algorithms as the initial population, thus achieving a reduction in the number of populations and a faster convergence[8].

We choose to use a small fully-connected neural network as the fitness function, which means implying a small training using the feature subset represented by each individual and the final accuracy as the fitness. it is more computationally intensive than both of the above approaches, but allows a more direct representation of the individual's quality, and also facilitates writing the leave-one-participant-out method[2]. In order to utilize the results of the MIC algorithm in the experiments, 40% of the individuals in the initial population are the encoded subset selected by MIC. The pesudocode is shown in algorithm 1.

```
Algorithm 1: Genetic algorithm
  Result: the individual with highest fitness
  initialization:
  set mutation_rate, crossover_rate, population_size, n_iteration, c_fitness
  pop1 = MIC, pop2 = random_population, P = pop1+pop2
  // Means 40 identical individuals by MIC and 60 random individuals
  fitness = Net(individual)
  // feed-forward neural network to evaluate fitness
  while n<n_iteration and fitness(f)<c_fitness do
      F = fitness(P);
      P = cxOnePoint();
      P = mutFlipBit();
      P = selTournament();
      p = selBest(P)
  end
  return p
```

2.3 Casper algorithm

The Casper algorithm is a modified cascade-correlation (Cascor) algorithm. Compared to a fully connected network, Cascor offers the advantage of no need to define network structure and fast convergence [9]. Cascor first builds a minimal topology, which means that the input nodes and output nodes are directly connected. After a certain time, Casper will add a new hidden neuron.

$$timeperiod = P * N + 15$$

This time threshold is defined as above, where P is a custom constant and N is the the number of hidden neurons already joined. The newly added hidden neuron is connected to all the previous neurons. The network will try to maximize the correlation between the residual error and the newly added neuron [10].

One drawback of the Cascor algorithm is that the addition of a new neuron ’freezes’ the previous weight, which means that only the weight connected with the latest addition has a non-zero learning rate. Later added nodes can only be trained with the interference of previous nodes. Casper uses the RPORP algorithm to achieve that all weights can be updated, as shown in the following figure[10]. As shown in figure2, Casper algorithm divides the weight into three parts, the first part is the weight connected to the newly added node, the second part is the weight connected to the output node of the newly added node

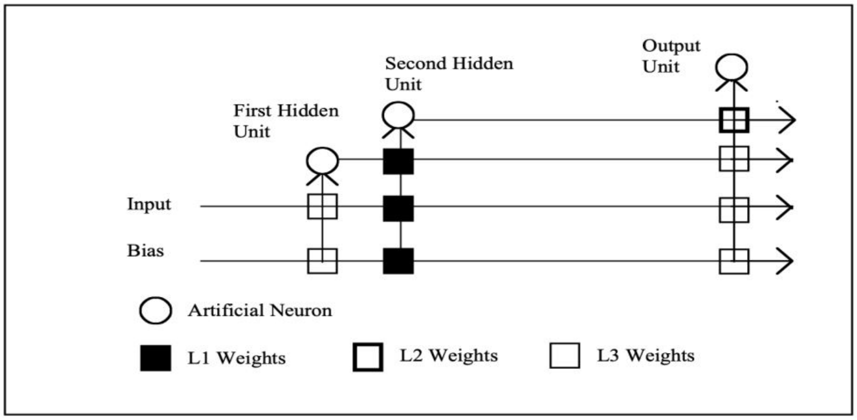

Fig. 2. Casper structure

and the output node, and the rest is the third part [10] their learning rate is L1 ⩾ L2 >L3. This allows the algorithm to both converge quickly after adding a new hidden neuron and to gradually correct the effects of the previous neurons. To improve the generalization ability, Casper algorithm also uses the simulated annealing (SA) method to make the learning rate gradually decrease[11].

$$\frac{\partial E}{\partial w_{ij}} = \frac{\partial E}{\partial w_{ij}} - k * sign(W_{ij} * W_{ij}^2 * 2^{-0.01*Hepoch})$$

The formula for SA is shown above, where k is a constant that can be set, sign means the positive or negative of its operand and Hepoch is the number of epochs that the neural network has been trained at the time the formula is executed.

The pesudocode of Casper is shown in algorithm 2.

Algorithm 2: Casper algorithm

```
Result: the individual with highest fitness
initialization:
set L1,L2,L3, n_iteration, n_checkpoint, max_hidden
net = Net()
X = features, Y = labels
D1,D2,D3 = dict()
while n<n_iteration do
    outputs = net(X)
    loss = criterion(outputs, Y)
    loss.backward()
    RMSprop()
    if n==checkpoint and n_hidden < max_hidden then
        net = Net.new_hidden(net);
        // Update D1, D2, D3 for storing all new connections
        para = Net.update_lr(net);
        // Assigning L1, L2, L3 learning rate to new connections
    end
end
```

The genetic algorithm's hyper-parameters are selected as follows: length of individual = 39, number of random individual = 12; number of MIC individual = 8; crossover rate = 0.8; mutation rate = 0.1; iteration number = 50. Note that the number of individuals and the number of iterations are small due to the limitation of the computing environment. This causes the disadvantage that the algorithm is difficult to converge (see figure3). To alleviate this situation, the mutation rate is also set low. The fitness function is a 3 layer feed-forward neural network which input size is the number of selected features, hidden size is 25. For Casper algorithm L1 = 0.2, L2 = 0.005, L3 = 0.001, max hidden neuron = 15. These parameters performed good on benchmark data[10], so I directly followed them. All the feed-forward neural networks mentioned above, including the fitness function and the baseline compared with Casper, use Adam optimizer and Leaky-ReLu activation function. The Casper algorithm use Leaky-Relu activation function and RMSprop optimizer as the paper described[10].

3 Experiment results and discussion

3.1 Evaluation

All the experiments are done on my personal laptop, with processor: Intel64 Family 6 Model 142 Stepping 10 GenuineIntel 1792 Mhz. The experimental part is divided into two parts as follows: 1. feature selection using genetic algorithm based on MIC algorithm to measure whether it can optimize the results of MIC algorithm. 2. comparison between Casper algorithm and feed-forward neural network using a subset of features after feature selection. For better evaluation, we use leave-one-out cross validation like [2]. Which means data from the same person will not appear in both training dataset and test dataset. This technique will be used by both feed-forward neural network and Casper algorithm for fair comparison, and the fitness function in genetic algorithm will also use this method.

3.2 Feature selection based on genetic algorithm

Since the fitness function is a neural network, this experiment is computationally expensive. Graph of the change of maximum fitness and generation is shown in figure 3. Because the results of the neural network depend on the initial value of weight and the training data used for each individual are different subsets sampled from the total training data, the final fitness has a certain amount of randomness. From figure3, we can see that this selection algorithm converges slowly.

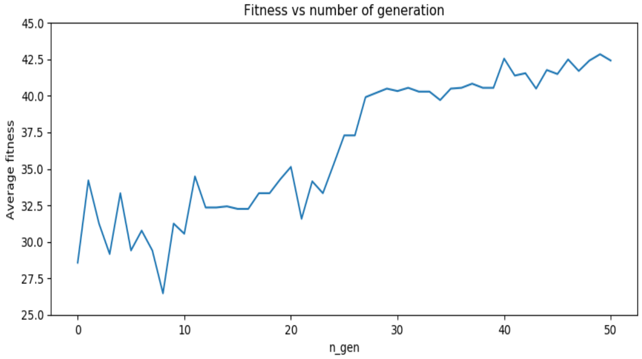

Fig. 3. Fitness changes of generations

3.3 Comparision between casper and feed-forward neural network

The control group is a three-layer fully connected network with 50 hidden nodes. The casper algorithm are implemented based on chehao2628's casper algorithm[12]. The maximum epoch is 500 and the maximum possible hidden neurons for casper and cascor algorithms is 15. The subsets of features selected by three different levels of feature selection strategies (raw data, MIC algorithm, and genetic algorithm) are used for comparison.The performance shown below:

Table 1. Performance for four class classification

Method	training accuracy	test accuracy	running time(second)	epoch used
Fully-connected(no feature selection)	1.00	0.306	4.525	50
Casper(no feature selection)	0.524	0.285	124.60	500
Fully-connected(MIC)	1.00	0.384	4.541	50
Casper(MIC)	0.701	0.401	98.05	500
Fully-connected(GA)	1.00	0.391	3.541	50
Casper(GA)	0.671	0.387	54.05	500

From Table 1, we can see that all the three networks are not very accurate. The best performance was achieved by Casper using MIC feature selection strategy, which reached 40%. From the results of the two different algorithms, the performance of GA and MIC is similar. However, GA uses fewer features, so it has relative advantage in running speed. Due to the small amount of data, the fully connected network is easily overfitted. The final result is highly dependent on the initial value. In my many experiments, the test set accuracy of the fully connected network fluctuates between 20% and 38%. The Casper algorithm, on the other hand, cannot converge before reaching the maximum number of epochs. So I think 4 class classification may be too complicated for this database, which leads to the following 2 class classification test.

Table 2. Performance for two class classification

Method	training accuracy	test accuracy	running time(second)	epoch used
Fully-connected(no feature selection)	1.00	0.591	4.45	50
Casper(no feature selection)	0.918	0.731	129.87	500
Fully-connected(MIC)	1.00	0.641	4.421	50
Casper(MIC)	0.904	0.814	89.23	500
Fully-connected(GA)	1.00	0.681	3.121	50
Casper(GA)	0.862	0.802	50.04	500

Under the premise that the accuracy of the 4 class classification is not satisfactory, Table2 shows the ability of these neural networks to distinguish between low level depression (original class 0 and class 1) and high level depression (original class 2 and class 3). All the three networks showed good accuracy in this case. The Casper algorithm converged successfully. The best performance was achieved by Casper using MIC feature selection strategy with 81.4%, but the performance of GA and MIC was still similar. This result illustrates that although detailed diagnosis still needs to be performed by a physician, current neural networks are competent to provide auxiliary diagnostic information.

In terms of computing time the Casper algorithm is significantly slower than the fully connected network. This is not consistent with the algorithm designer's claim that it is about ten times faster[9]. This may be because modern computers are better suited to run fully connected network, or it may be because the algorithm I used to assign the learning rate is not as efficient as the inbuilt update algorithm in pytorch. Feature selection effectively improves the running speed of the algorithms and does not affect the accuracy. From the experiments it can be seen that the genetic algorithm does not perform well in improving the accuracy, but it can improve the training speed by reducing the size of the feature subset. However, if a neural network is used as the fitness function, the genetic algorithm itself is computationally expensive, and the extra time spent is much larger than the training speed improvement. Due to the excessive time required on a personal computer, the same size populations as described in [2] are not used in this paper. However, for multiple individuals in a population, the computation of fitness is clearly parallelizable, so genetic algorithms may still be valuable in situations where parallel computation is supported.

4 Conclusion and Future works

In this paper we compare the performance of Casper algorithm and the fully connected network based on a small training set, under two feature selection approaches, exploring their ability to identify the degree of depression. Due to the limitations mentioned in the previous sections, all the two methods do not have high accuracy in the 4 class problem. However they still shows relatively good accuracy for a simple binary classification problems (80%). When dealing with small dataset, cascade correlation algorithms are more difficult to converge. The difference between the actual runtime and the theoretical runtime[9] also indicates that there is room for optimization of the algorithm used in this paper. Due to the absence of a parallel computing environment, the genetic algorithm is not adequately studied in this paper. Both the number of individual and iterations are too small, so that the algorithm does not converge well. For this reason, a similar approach as in [8], i.e., using the results of several conventional feature selections as the initial population to reduce the randomness, may achieve a higher accuracy. Possible future research directions include improving the running speed of Casper and trying to continue testing Casper's performance with larger datasets. For the genetic algorithm, the first step is to use a better fitness function or parallel method to increase the running speed, and the fitness function can be modified with a function related to the subset size, thus further reducing the number of features selected with similar accuracy. In addition to directly reducing the running time of the algorithm, a smaller number of features can also be used to try to obtain a higher accuracy by applying some dimensionality enhancement techniques such as support vector machine (SVM).

References

1. K. Hawton, C. C. i Comabella, C. Haw, and K. Saunders, "Risk factors for suicide in individuals with depression: a systematic review," *Journal of affective disorders*, vol. 147, no. 1-3, pp. 17–28, 2013.
2. X. Zhu, T. Gedeon, S. Caldwell, and R. Jones, "Detecting emotional reactions to videos of depression," in *2019 IEEE 23rd International Conference on Intelligent Engineering Systems (INES)*, pp. 000147–000152, IEEE, 2019.
3. M. Vermani, M. Marcus, and M. A. Katzman, "Rates of detection of mood and anxiety disorders in primary care: a descriptive, cross-sectional study," *The primary care companion to CNS disorders*, vol. 13, no. 2, 2011.
4. F. Amato, A. López, E. M. Peña-Méndez, P. Vaňhara, A. Hampl, and J. Havel, "Artificial neural networks in medical diagnosis," 2013.
5. J. B. Kinney and G. S. Atwal, "Equitability, mutual information, and the maximal information coefficient," *Proceedings of the National Academy of Sciences*, vol. 111, no. 9, pp. 3354–3359, 2014.
6. M. Mitchell, "An introduction to genetic algorithms massachusetts," 1998.
7. O. H. Babatunde, L. Armstrong, J. Leng, and D. Diepeveen, "A genetic algorithm-based feature selection," 2014.
8. F. Tan, X. Fu, Y. Zhang, and A. G. Bourgeois, "A genetic algorithm-based method for feature subset selection," *Soft Computing*, vol. 12, no. 2, pp. 111–120, 2008.
9. S. E. Fahlman and C. Lebiere, "The cascade-correlation learning architecture," tech. rep., CARNEGIE-MELLON UNIV PITTSBURGH PA SCHOOL OF COMPUTER SCIENCE, 1990.
10. N. K. Treadgold and T. D. Gedeon, "A cascade network algorithm employing progressive rprop," in *International Work-Conference on Artificial Neural Networks*, pp. 733–742, Springer, 1997.
11. N. Treadgold and T. Gedeon, "The sarprop algorithm, a simulated annealing enhancement to resilient back propagation," in *Proceedings International Panel Conference on Soft and Intelligent Computing*, pp. 293–298, 1996.
12. chehao2628, "Improvedcasper," 2020.