Fantastic optimization of oil well data prediction model using Gan GA and Pruning

Ziyu Cui[1]

{Australian National University CECS master of computing, u7004551}@anu.edu.au

Abstract. In this paper, the linear regression model is used to predict the phi and logK of the oil reservoir in oil wells. Based on the regression model, I used a variety of techniques to optimize the underlying model. I use Gan to extend the model. I use a genetic algorithm to adjust the parameters of the training model. I use pruning techniques to reduce the parameters of the model. Three techniques effectively reduce the accuracy of the model. And it improves the generalization ability of the model

Keywords: linear regression, neural network, GAN, pruning, genetic algorithm

1 Introduction

Permeability and porosity of oil reservoir are important parameters for reservoir evaluation. The two parameters have some correlation with other available parameters in oil well exploration. Therefore, it is an efficient method to predict the porosity and permeability of the oil reservoir by using other parameters. Moreover, in the oil well development industry, This method has also been widely used. Many research institutions use a variety of mathematical models to model oil reservoir data. For example, Kozeny-Carmen theory relates permeability to porosity and the specific area of a porous rock with pores treated as an idealize bundle of capillary tubes. However, it is challenging to predict permeability and porosity accurately only by a statistical method and mathematical formula. This is mainly because the linear relationship between permeability and porosity of other observation parameters is relatively complex. Furthermore, in the process of exploration, there will be a lot of errors, which will make some noise in the data set, and ultimately affect the construction of the model.

Machine learning has flourished in recent decades. Based on the theory of linear regression, using backpropagation makes the automatic training of multi-layer neural network easier. In the interdisciplinary, linear regression model has been widely used, and in the continuous improvement of researchers, the model has good stability. By using a linear regression model, it is easy to find out the

complicated relationship between input and output. Furthermore, this technique does not require researchers to have good knowledge in different fields.

Because the data of oil well is not as significant as other machine learning tasks, the amount of data of single oil well is minimal. For such a small data set, learning neural network is a challenging job. Only using simple regression model can not meet the training accuracy. More deep learning techniques are needed to help improve the accuracy of neural networks.

2 Related works

Ian Goodfellow proposed generative Adversarial Networks(GAN) in 2014. GAN can be used for data generation. When our data volume is small, we can expand our data set by generating a small amount of data. Moreover, GAN is a relatively ideal method. The Gan network usually consists of two sub-networks, one is a generative model, the other is discrimination model. The generative model transforms a random segment of noise into the data we need. Identify the data generated by the model pair and the real data. The two models form an adversarial relationship. This confrontation process leads to the convergence of the whole model.

Genetic algorithm is a computational model simulating the natural selection and genetic mechanism of Darwinian biological evolution. It is a method to search the optimal solution by simulating the natural evolution.

Its main characteristics are direct operation on the structure object, without the limitation of derivation and function continuity, inherent, implicit parallelism and better global optimization ability, using probabilistic optimization method, the search space can be obtained and guided automatically without specific rules, and the search direction can be adjusted.

The genetic algorithm takes all individuals in a population as objects and uses randomization technology to guide the efficient search of coded parameter space. Among them, selection, crossover and mutation constitute the genetic operation of the genetic algorithm; parameter coding, the setting of the initial population, design of fitness function, design of genetic operation and setting of control parameters constitute the core content of genetic algorithm.

The genetic algorithm can selfly iterate, let its system of things for the survival of the fittest natural selection, keep the good, the next thing will be excluded. The essence of the genetic algorithm is the survival of the fittest, the selection of the best individuals generally used to find the best solution. One of the advantages of genetic algorithm is that it can prevent the model from falling into the optimal local solution.

3 Method

3.1 Generative Adversarial Networks

Due to the small amount of oil well data, Gan is used to expanding the data set in research. In GAN, I use the complete training set as the training of GAN. The input contains ten parameters. This includes eight input variables and two prediction variables in the regression network. Gan network can not only increase the number of training data but also fit the data in a higher dimension. This can help to improve the generalization ability in the following predictive neural network.

3.1.1 Generator

In generator, a random noise data with a length of 10 is used to generate the oil well data we need. The diversity of output values can be increased by using random input variables. The generator uses the neural network of two hidden layers, and the input is random noise with a length of 10. The first layer and the second layer contain 64 units. The output is 10 lengths of data. The hidden layer uses well as the activation function. The output uses sigmoid as the activation function.

3.12 Discriminator

The discriminator consists of a two-layer fully connected neural network. Enter well data of length 10. The first and second layers have 64 units. The output is a value. Use selu as the activation function in the hidden layer. The output uses sigmoid as the activation function.

The loss function of the discriminator is a cross-entropy function. The closer the result is to 1, the higher the probability that the data is real. The closer the result is to 0, the higher the probability that the data is false.

There are two training steps in the discriminator. Firstly, the discriminator needs to be trained to recognize the true and false data. Using the real test set and the false data generated by the noise to distinguish, improve the ability to distinguish the true and false data. Secondly, we need to train the generator by targeting the real data.

3.13 Training

In training, the discriminator and generator are alternately trained to form a kind of confrontation training. The loss of discriminator will increase because of the training of the generator. However, the loss trend of the two training is declining.

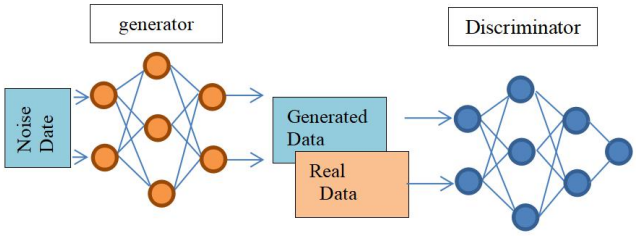

Fig.1 Gan model structure

3.2 Dataset

Dataset is from “A clustering assisted method for fuzzy rule extraction and pattern classification. In ICONIP'99. ANZIIS'99 & ANNES'99 & ACNN'99. 6th International Conference on Neural Information Processing”.

The data in the data set has been normalized. Data set is divided into a training set and test set.

3.2.1 cross-validation

Because the data set is small, I use the cross-validation method to divide the data set into N parts on average, select N-1 of them as the training set, and the remaining one as the validation set. Cross-validation can make full use of limited data to find appropriate model parameters and prevent overfitting.

I divide the training data into five parts, one as the validation set, four as the training set, and conduct cross-validation.

3.3 Regression Model

In this research, a predictive neural network is a right choice. The predicted result is one or more continuous variables. Oil reservoir data of oil well mainly includes GR, rdev, rmev, rxo, RhoB, nphi, PEF and DT. The predicted data are phi and logK. Therefore, the neural network is a two-layer hidden layer neural network with eight inputs and two outputs, as shown inFig.1.

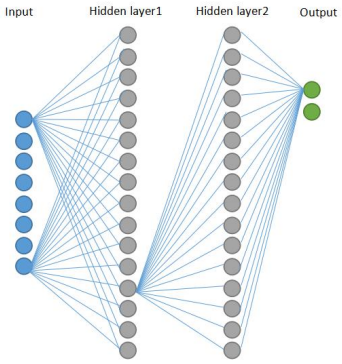

Fig.2 The Structure of regression model

The formula for the model is as follows:

$$f(x) = h(b_2 + g(b_1 + g(b_0 + w_0 \cdot x) \cdot w_1) \cdot w_2)$$

The activation function is a kind of nonlinear function. The expression ability of the linear model is not enough. The activation function increases the nonlinearity of the neural network model and improves the expression ability of the neural network model.

The loss function is MSE, which is the most commonly used activation function in regression tasks. Its formula is:

$$\mathrm{MSE} = \frac{1}{\mathrm{n}_p} \sum_{k=1}^{n_p} (f(x_k) - y_k)^2$$

The loss function is a way to measure the difference between the predicted value of the output of the neural network and the actual value. The gradient of each parameter in the model can be calculated by a loss function, and the whole model can be updated by gradient descent method.

3.4 Genetic Algorithm

3.4.1 Encoding

In the genetic algorithm, I use all the parameters in the model as chromosomes. Each individual contains six chromosomes, which are weight and bias of two input layers and one output layer, respectively. Each float parameter in weight and bias is a gene. This is a float encoding. A small random deviation is used as the variation of each gene. This encoding has the following benefits for binary encoding. 1. The encoded values are all valid data. 2. The high precision can be applied to the continuous variable problem, avoiding the Hemingway cliff problem. 3. Reduce computational complexity and improve efficiency.

3.4.2 Target Function and Selection

The chromosomes of each individual are divided into parameters of each layer of the neural network. Reassign these parameters to my prediction model. I use the training set to predict and evaluate the loss of the model. The fitness is as follows.

$$fitness = 1 - loss, 0 < loss < 1$$
$$fitness = 0, loss > 1$$

In this way, the smaller the loss, the higher the probability of being retained when generating the next generation.

The selection is based on the fitness evaluation of individuals in the population, and the genes of individuals with high fitness value are easy to inherit to the next generation. I use Boltzmann selection—this selection based on the thermodynamical principles of simulated annealing. Selection Probability is as follows.

$$\varphi(x_i(t)) = \frac{1}{1 + e^{f(x_i(t))/T(t)}}$$

T(t) is the temperature parameter. A sizeable initial value ensures that all individuals have an equal probability of being selected. As T(t) becomes smaller, the selection focuses more on good individuals.

3.4.3 Crossover

In different individuals of crossover, alleles of the same chromosome have a certain probability of cross-exchange. In research, I set the probability to 80%. The algorithm uses uniform crossover. Each gene has a certain probability and allele to crossover.

3.4.4 Mutation

In the process of breeding, some mutation may occur. Only random changes of a gene are considered here. Change the gene value of some sites in the gene chain of the parent individual. Moreover, form the operation of new individuals with a certain probability. Mutation operation is also one of the core algorithms of the genetic algorithm, aiming to jump out of the scope of local search and embody the idea of local search.

The mutation results in a small shift in each gene. This offset has two directions, positive and negative.

$$\mathrm{DNA} = \mathrm{DNA} + \mathrm{mutation\_factor}$$

3.5 Training

Training consists of two steps. First, the training data is generated by GAN. GAN helps me generate 200 new data. Second, train the data, use a regression model to train the training set, and use the test set to evaluate the data. After 1000times of training, the model has not reached the optimal solution. Genetic algorithm is used to further fine-tune the parameters of the model. After 7000 times of training, I import the parameters of the model into the regression model for the second training. In order to improve the accuracy of the model.

3.6 Optimization after training

Pruning the model can reduce the number of neurons in the model and improve the calculation efficiency of the neural network.

In neural networks, some neurons may not play a role, and some hidden units will play the same role. Some hidden units play the opposite role. The two hidden units are consistent with each other so that the expression of the two hidden units to the next layer is 0. These types of hidden units will not play a useful role in the neural network but will waste the calculation of the model, resulting in the low efficiency of the neural network, so we need a reasonable means to prune the hidden layer neurons.

There are many kinds of pruning techniques, but in this research, only two kinds of hidden units are pruned. They were pruning the hidden units with the same function and units with the different function. Each unit is associated with the weight of the previous link, which constitutes the vector of each hidden units. So the relationship between each neuron can be known by comparing the vector of each hidden unit. We compare the angles between the two hidden units. If the angle between the two hidden units is less than 15°, the information expressed by the two units is highly similar in this layer. So we only merge the vectors of two cells and delete one cell. When the angle between the two vectors is more significant than 165°, the expression value of the two hidden units to the next layer is 0, which makes the two hidden units invalid. In this case, we can delete these two neurons.

4 Experiment result

We can see that Gan is in a state of confrontation in the early two models. Moreover, the overall trend gradually converges. The final loss tends to 0. It is proved that Gan can generate well data well

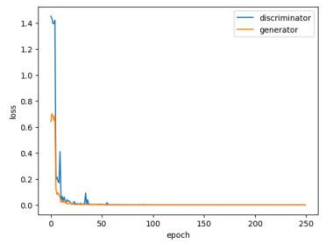

Fig.2 Process of training Gan, loss reduction of discriminator and generator

After using the data generated by Gan, the loss of training set decreased a lot, from about 0.39 to about 0.22. The loss of the test set also decreased. This proves that the performance of the model has been improved after adding GAN.

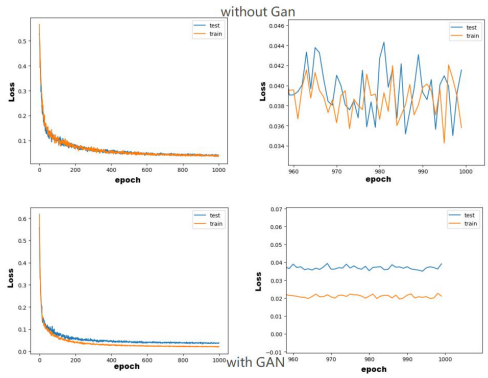

Fig.3 Comparison before and after using Gan

Using the genetic algorithm, the loss of the model can be further reduced, and the result is as shown in the figure.

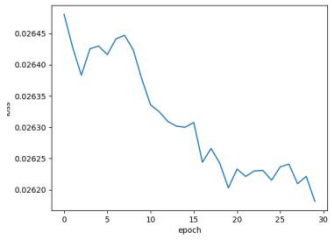

Fig.4 The change of loss in the training process of genetic algorithm

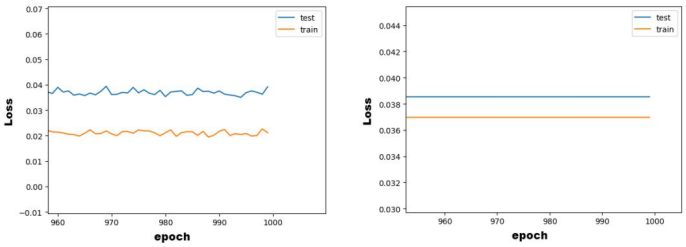

Fig.5 Comparison before and after using GA

5 Discussion

After the optimization of the data set and neural network, the loss of neural network decreased significantly. It is not easy. Because the data set is so small, I may not be able to make any other better improvements to this model.

I think we have done enough research on the good data set. I have improved my accuracy considerably from the original. However, because the data set is too small is still a major problem. I do not think it's necessary to improve this model if we cannot get a more extensive data set. It is impossible to improve accuracy. However, I think if we still want to dig into the value of this data, we need to rely on traditional algorithms rather than deep learning. We need to work with more professional people, such as geographers and chemists, to do a professional analysis of the penetration rate of the well. Find out the exact relationship between different variables in the well. Then the prediction value can be obtained by direct mathematical calculation, which is the real need of high-quality algorithm for the shale gas industry.

Reference

1 Wong, P.M., Jang, M., Cho, S. and Gedeon, T.D., 2000. Multiple permeability predictions using an observational learning algorithm. Computers & Geosciences, 26(8), pp.907-913.

2 Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. and Bengio, Y., 2014. Generative adversarial nets. *In Advances in neural information processing systems* (pp. 2672-2680).

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

4 Wang, X., Man, Z., You, M. and Shen, C., 2017. Adversarial generation of training examples: applications to moving vehicle license plate recognition. *arXiv preprint arXiv*:1707.03124.