Sampling Method Analysis and Quality Evaluation Strategy for Remote Sensing Big Data

Dang Yu[1], Zhang Jixian[1,*], Zhang Pengcheng[1], Luo Fujun[1], Bai Jin[1]

[1] National Quality Inspection and Testing Centre for Surveying and Mapping Products,

KEY WORDS: Natural Resources, Remote sensing, Big Data, Quality Evaluation, Sequential sampling

ABSTRACT:

Under the background of the increasingly unified management of natural resources, remote sensing big-data will become the main data source to support a number of major projects. How to sample the natural resources results efficiently and reliably in the process of quality evaluation is always a research hotspot when it comes to the natural resources results involving remote sensing big-data. A sequential quality evaluation model based on root mean square error (RMSprop) optimization algorithm is constructed by theoretical analysis with an numerical experiments to validate the effectiveness of this method.

1. INTRODUCTION

In the existing sampling process of Surveying and mapping product quality evaluation, the design goal of sampling is to minimize the overall risk. A risk minimization model is designed to extract a small number of samples to achieve the overall quality evaluation. This paper puts forward the idea of building a quality model of remote sensing large data results, and realizes the evaluation of the quality of remote sensing results by mining the data characteristics of the results. An important basis of this method is to verify the validity of multiple sequential sampling with small sample size. Therefore, this paper will mainly study the validity, advantages and experimental validation of this sampling method.

The production process of remote sensing data is generally stable. Assuming that the quality model of data results conforms to the basic independent and identical distribution, similar validity can be achieved by extracting small sample size data and increasing the number of sampling times. More sampling will produce more noise in quality evaluation results, so the RMSprop optimization algorithm which is suitable for this sampling method can largely suppress the noise caused by multi-batch small data sampling. To achieve effective evaluation through multiple sampling of small sample size and the algorithm adapted to the quality evaluation model, and at the same time to reduce the manual interpretation brought about by mass data quality inspection.

2. STOCHASTIC GRADIENT DESCENT(SGD) AND MULTIPLE SEQUENTIAL SAMPLING WITH SMALL SAMPLE SIZE

Stochastic gradient descent (SGD) is a widely used optimization algorithm in the field of machine learning. Gradient descent can be expressed as the process of minimizing the risk function of loss function j (θ) in a specific dimension space. Taking linear regression function as an example, the loss function j (θ) around the objective function h (θ) is constructed, and the fitting of the objective function is realized by iterating θ. Assuming there are m samples and N features, the objective function and loss function are shown as follows (1), (2).

$$h(\theta)=\sum_{j=0}^{n}\theta_j x_j$$

$$J(\theta)=\frac{1}{2m}\sum_{i=1}^{m}(y^i-h_\theta(x^i))^2 \quad (1),(2)$$

The gradient of partial derivative of θ is obtained by calculating the loss function j (θ), and the iteration of θ is achieved by negative gradient, and the objective function is fitted as shown in equation (3), (4). Since gradient descent requires calculation of each sample and each feature, under the assumption that there are m samples and n features, the computational cost of iteration is $\mathbf{m} \times \mathbf{n}^2$.

$$\frac{\partial J(\theta)}{\partial \theta_j}=-\frac{1}{m}\sum_{i=1}^{m}(y^i-h_\theta(x^i))x_j^i$$

$$\theta_j^{'} = \theta_j + \frac{1}{m}\sum_{i=1}^{m}(y^i - h_\theta(x^i))x_j^i \quad (3),(4)$$

Stochastic gradient descent is an improved algorithm based on gradient descent. A new risk minimization function is constructed by rewriting the loss function. The loss function j（θ）can be rewritten to a single sample (5):

$$J(\theta) = \frac{1}{m}\sum_{i=1}^{m}\frac{1}{2}(y^i - h_\theta(x^i))^2 = \frac{1}{m}\sum_{i=1}^{m}\cos t(\theta,(x^i,y^i))$$

$$\cos t(\theta,(x^i,y^i)) = \frac{1}{2}(y^i - h_\theta(x^i))^2 \quad (5)$$

The gradient updating of each θ can be rewritten to equation (6):

$$\theta_j^{'} = \theta_j + (y^i - h_\theta(x^i))x_j^i \quad (6)$$

Comparing the two optimization methods, the calculation amount of random gradient descent iteration is n^2. When the data amount m is large, the random gradient descent has advantages in iteration speed and calculation amount. But because of the smaller sample size, the noise of random gradient descent is larger. In large data background, it can be overlapped at a faster speed. The random gradient descent is generally better than the calculation of all the data because there are more rounds.

In this paper, the idea of random gradient descent of small sample set is applied to the sampling process of quality evaluation of remote sensing results for large sample size. At present, large-scale surveying and mapping projects are generally accompanied by large-scale remote sensing image results, which challenges the current sampling methods and manual inspection. By putting forward an evaluation method for the quality model of remote sensing results, it can be assumed that there exists an objective multi-dimensional and multi-parameter quality feature model for each remote sensing result. If the model can be effectively fitted by the method in this paper, the sampling process of the result can be incorporated into the evaluation system. Through the manual inspection of the small sample size data extracted from the results many times and the application of the evaluation results as intermediate data in the deep neural network model applicable to the quality evaluation of Surveying and mapping results, the effective sampling and evaluation of large-scale quality results can be realized, that is, a small sample size multiple sequential sampling evaluation method.

3. CHARACTERISTICS AND ADVANTAGES OF MULTIPLE SEQUENTIAL SAMPLING WITH SMALL SAMPLE SIZE

Sequential sampling adopts the strategy of uncertain population sampling quantity, and decides the next sampling mode according to the results of the preceding sampling. Compared with the fixed sampling, the sequential sampling method can determine whether the sampling is adequate or not by combining the sampling process to obtain more stable results.

Under the background of remote sensing big data, the scale of data is getting larger and larger. Fixed sampling is facing unprecedented challenges. Because the large-scale projects that produce large-scale results data usually has long production cycle, many production units and complex quality characteristics, in the process of inspection for such results, in order to meet the needs of the project, the results inspection work will generally be carried out in stages, but because of many uncertainties in the actual project implementation process, the inspection work will be carried out in stages. Often occurs in the short term need for a large number of manual inspections. By using the method of multiple sequential sampling with small sample size, it is easier for each batch to form inspection data set, which has positive significance for timely submission of inspection samples. At the same time, this method reduces the amount of data needed to be inspected, thus shortening the period of result test which means faster iteration and more time-consuming.

Another advantage of multiple sequential sampling with small sample size is that it does not establish a direct relationship between the quality of the samples and the quality of the overall results. Each iteration is a process of minimizing the risk function of the quality model, thus reducing the sampling risk of each batch of samples. An important premise of adopting this method is that the final conclusion obtained by multiple small sample size sampling and small batch large sample size sampling is similar, that is, this kind of sampling method has similar reliability with the existing sampling methods. The validity of this kind of sampling method will be analysed with a numerical experiment.

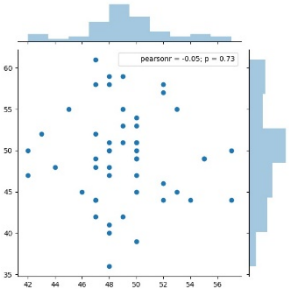
4. EXPERIMENTAL AND ANALYSIS

The experimental verification of the method in this paper simulates the corresponding quality characteristics and corresponding quality elements of remote sensing data by generating a large number of random arrays which conform to a probability distribution characteristic. In order to verify the robustness of the evaluation model to noise, some random noise is added to the experimental data. Sampling noise of small data sets and noise perturbation of multi-dimensional arrays simulate additional noise perturbation caused by a small number of sampling noise and accidental factors on the quality of a small number of results. After obtaining the data set, the validity of this method is verified by comparing the evaluation results of this method with whether it can extract the hidden features in the data.

4.1 Generating experimental data sets

The numerical simulation needs a high-dimensional array with non-linear mapping and fixed features, which contains fixed features to characterize the quality characteristics of remote sensing big data. At the same time, in order to verify the reliability of the method, it is necessary to disturb the data and simulate the complexity of the real situation.

By generating two groups of random arrays with 50 dimensions, one group is noise perturbation with normal distribution, the other group is normal distribution with mean value of 50 and variance of 3, 5, 8 and 12 as four gradient data. The two arrays are fitted to produce a distribution density image by smoothing the two-dimensional kernel density estimation, as shown in the following figure:



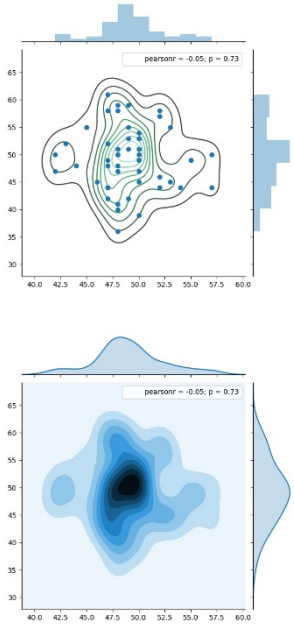


Figure 1. Data generation process schema

In the process of generating a pair of data randomly, two multi-dimensional arrays will generate 100 discrete points on the basis of matching characteristics. The data coupling is realized by KDE mode for the next experiment. One of the final data arrays has four gradient variances, resulting in 800 image data to simulate data with four different quality characteristics, as shown in the following figure:

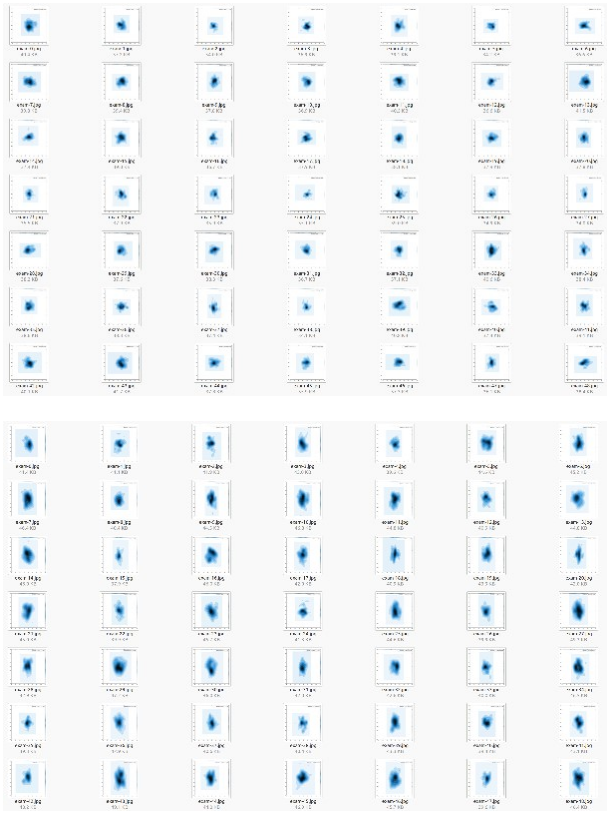


Figure 2. Data set schema

4.2 Multiple Learning with Small Sample Size by Neural Network

A 19-layer convolution neural network is used as a learning model for each batch of samples to realize feature mining for small data volume with multiple batches. The data sets with 800 data sets were sampled with 4% small sample size of 32 pieces per batch and recycled 2000 times. The conclusion of the validity of sampling was obtained by analysing the accuracy of data classification improved with the increase of rounds. The risk function decline process of the training process is shown in the following figure:

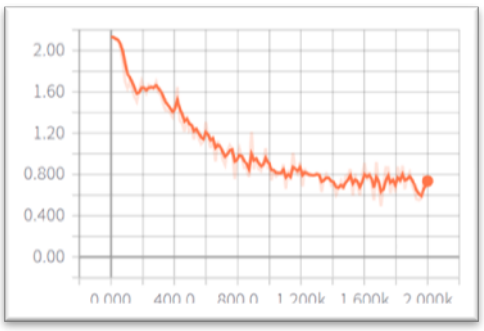


Figure 3. The Decline Process of Risk Function with the Increase of Rounds

The validation set is used to validate the model. The model can classify the four gradient samples with 92.3% accuracy. Since the data set contains 800 pieces of data and 32 pieces of data are extracted at a time, feature learning equivalent to a complete sample amount can be obtained at 25 times of sampling, which is generally called an epoch in the training of neural networks. 2000 rounds of learning is equivalent to 80-epoch, that is, 80 rounds of complete learning for all data. The risk function converges to a relatively stable minimum after 1400 sampling, i.e. 56-epoch.

In this paper, a batch of 8, 16 and 32 datas were sampled in a similar experimental environment, and the comparative experiments were carried out in 1%, 2% and 4% cases respectively. The experimental results are as follows:

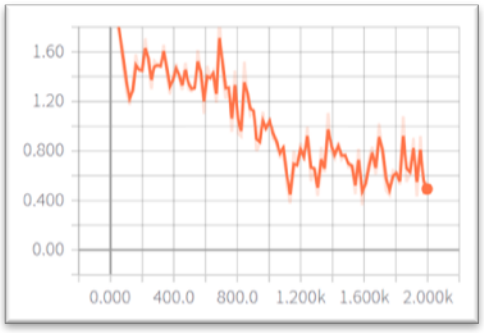


Figure 4. 1% sample size results

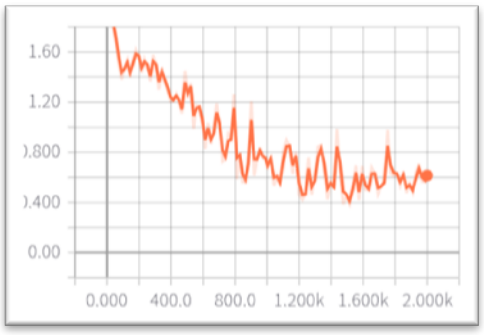


Figure 5. 2% sample size results

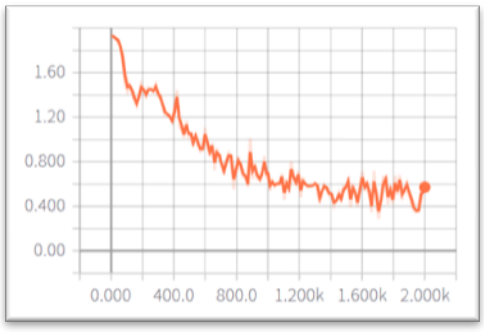


Figure 6. 4% sample size results

The experimental results show that the smaller the sample

size of single extraction, the greater the noise disturbance of the whole curve. The final results of the three experiments tend to converge, but the final accuracy of 1% sample size is 82.3%, the final accuracy of 2% sample size is 89.9%, which is lower than 92.3% of the 4% sample size. In other comparative experiments with less than 1% sample size, the final convergence will not be possible.

4.3 Experimental analysis

The experimental results show that after 1400 sampling times, this method effectively obtains four different variance gradients in highly coupled data by using 4% small sample size extraction method, and distinguishes the gradients of any sample with 92.3% accuracy. That is to say, it realizes the mining of batch sample features.

The comparative experimental results also confirm the limitations of the small sample size multiple sampling method, that is, with the reduction of sample size, there will be greater noise in the process of feature mining. But through the deep learning neural network method and root mean square error optimization algorithm adopted in this paper, the noise in the process of small batch sampling can be suppressed in an acceptable range and the data features can be effectively mined.

Because the feature data generated in this experiment is still far from the actual remote sensing data in complexity, there are still some practical problems in the application of large data for remote sensing, such as regularization method needed for data mining of remote sensing data features, minimum sample size, effective convergence rounds and product characteristics. Establishment of reliable models and parameters. These are all problems that need to be solved in practical engineering. However, numerical experiments in this paper have explored the feasibility of these methods in theoretical level and achieved expected results.

5. CONCLUSIONS

In the existing engineering practice, due to the small scale of data and the limitation of production mode, it is not yet mature to evaluate the quality of remote sensing data by means of large data mining. Under the background of natural resources, the amount of remote sensing data will continue to increase dramatically in the future. The new sampling and quality evaluation model can bring shorter inspection term, faster iteration speed and reliable sampling method for quality evaluation in engineering. Combining the advantage of feature expression of large data with the background of large remote sensing data in the new era can form a new quality evaluation strategy adapted to the characteristics of large remote sensing data.

At the same time, in order to effectively combine the results of artificial interpretation and add them to the final evaluation model, it is necessary to conduct a larger-scale experimental study with specific types of remote sensing results in order to obtain the neural network model which can effectively form the evaluation system of the model and the pre-training model corresponding to each type of results, which still needs to be carried out. A lot of work can be obtained. This paper mainly carries out experiments to verify the validity of the intermediate process of sampling, and explores a new achievement evaluation system under the background of remote sensing big data.

ACKNOWLEDGEMENTS

This work is partly supported by National Science Foundation of China (grant No. 41671440)

REFERENCES

L. Zhang, L. Zhang, and V. Kumar, "Deep learning for Remote Sensing Data," IEEE Geoscience and Remote Sensing Magazine 4(June), 22–40 (2016).

J. Wan, D. Wang, S. C. H. Hoi, et al., "Deep learning for content-based image retrieval: A comprehensive study," in Proceedings of the 22nd ACM international conference on Multimedia, 157–166 (2014).

I. Arel, D. C. Rose, and T. P. Karnowski, "Deep Machine Learning - A New Frontier in Artificial Intelligence Research," IEEE Computational Intelligence Magazine 5(4), 13–18 (2010).

H. Deborah, N. Richard, and J. Y. Hardeberg, "A comprehensive evaluation of spectral distance functions and metrics for hyperspectral image processing," IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 8(6), 3224–3234 (2015).

P. Dollar, C. Wojek, B. Schiele, et al., "Pedestrian detection: An evaluation of the state of the art," IEEE transactions on pattern analysis and machine intelligence 34(4), 743–761 (2012).

M. Fauvel, Y. Tarabalka, J. A. Benediktsson, et al., "Advances in spectral-spatial classification of hyperspectral

images," Proceedings of the IEEE 101(3), 652–675 (2013).