Study of the Application of Disagreement-based Collaborative Regression in Log Interpretation

Yu-zhe ZHENG¹, Zhao-hui YE¹ and Cong-hui ZHANG²
¹Department of Automation, Tsinghua University, Beijing 100084, China
²China Oilfield Services Limited, Sanhe, Hebei 065201, China

Keywords: Semi-supervised learning, Disagreement-based, Collaborative regression, Log interpretation.

Abstract. In the field of log interpretation, it’s easy to acquire a lot of data, however, it requires cost to get the label information, thus the labeled samples are often not enough. The secondary interpretation of porosity is a typical application task with few labeled samples and a large number of unlabeled samples. Manual interpretation has the disadvantages of strong subjectivity and low accuracy. A disagreement-based co-training style semi-supervised regression algorithm was proposed as an alternative to the manual interpretation. Two kNN regressors with disagreement were employed. Each of them labels unlabeled samples with high confidence level for the other to improve regression estimates. The method was verified through the experiments which showed that the generalization performance of this semi-supervised model is better than the other supervised models in such cases.

Introduction

In the process of petroleum exploration[1], people use special instruments, such as acoustic wave, electricity, and radioactivity, to measure various parameters of the stratum at different depths in the well, and then analyze the parameters. This is called log interpretation. Because reservoir resources are generally distributed in underground interconnected pore cracks or rocks, prediction of porosity is important in the process of log interpretation. Prediction of porosity includes primary interpretation and secondary interpretation. Traditionally, people use the data acquired to calculate porosity in the process of primary interpretation based on response equation. But the secondary interpretation needs to be carried out through core analysis. Core analysis is the process of measuring the actual porosity in rock samples at some depth in the underground and then correcting the primary interpretation result by these samples. In actual production, the result of primary interpretation is often inaccurate, and core analysis also requires a lot of cost. Moreover, secondary interpretation through artificial methods is highly subjective and has high requirements for technicians.

In fact, the analysis of an well can be carried out with the aid of the existing historical data of other wells in the same area. Artificial intelligence technology can independently discover and learn rules from existing historical data and predict the output of new samples. Its way of processing data is completely different from traditional theory. Many scholars have applied artificial intelligence technology to log interpretation[2][3][4], but most of these are supervised methods. In log interpretation industry, usually the cost of getting the label information is high, so it’s often hard to get too many labeled samples. Therefore, the accuracy of these supervised learning methods is often not high.

Because core analysis can only be carried out at some depth, the secondary interpretation of porosity is a typical application where the data set consists of a small number of labeled data and a large number of unlabeled data. Experience shows that generally supervised methods tend to fall into overfitting for situations where labeled samples are scarce, while semi-supervised method can make use of unlabeled samples and perform better. At present, there are few studies of semi-supervised learning application in log interpretation industry. In this paper, we propose to apply a semi-supervised collaborative regression method to the secondary interpretation of porosity. The
experiments with the actual production data of China Oilfield Services Limited show that collaborative regression is superior to other methods in this application.

**Semi-supervised Learning**

With the development of modern information technology, it is usually easy to acquire a large number of unlabeled samples in many fields, but it takes some cost to get the label information[5]. The generalization performance of supervised learning method is limited by the number of labeled samples, and if only unsupervised learning is adopted, the value of labeled samples is wasted, while semi-supervised learning method can make use of both labeled and unlabeled samples[6].

At present, the most popular semi-supervised learning method is disagreement-based collaborative learning, which takes advantage of the differences between multiple classifiers or regressors to make use of unlabeled samples. It has the advantages of few assumptions, simple and effective learning methods, and wide application scope so it’s the mainstream algorithm in semi-supervised learning currently. Figure 1 is a schematic diagram of the disagreement-based collaborative learning[7]. For application of disagreement-based collaborative learning in regression problems, Zhou and Li[8] put forward that the authenticity of the sample labels estimated by the regressor can be determined by examining the confidence level of the samples and samples with high confidence level is consistent with the trend of regression. In this paper, a disagreement-based collaborative learning method is designed for the secondary interpretation of porosity, which will be detailed in the next section.

![Schematic diagram of the disagreement-based collaborative learning.](image)

**Secondary Interpretation of Porosity Based on Disagreement-based Collaborative Learning**

Traditional supervised methods only utilize labeled samples. This algorithm mainly solves the problem how to make use of unlabeled samples to improve the generalization performance under the circumstances where labeled samples is insufficient.

Let \( L = \{(x_1, y_1), (x_2, y_2), \ldots (x_n, y_n)\} \) denote the labeled sample set, and \( U = \{x'_1, x'_2, \ldots x'_n\} \) denote the unlabeled sample set. Disagreement-based collaborative regression utilizes set \( L \) and \( U \) to train a regressor \( f : X \rightarrow Y \). The process of the algorithm is designed as follows.

**Initialize the Data Set**

Randomly pick \( N_L \) samples from labeled data set to form set \( L \) used for training, and the remaining data is retained as test set. Then randomly pick \( N_U \) samples from unlabeled data set to form set \( U \).

**Configure Regressor**

In this paper, kNN regressor, which is simple but effective, is used as the base learner. The configuration of the regressor includes determining the neighbor number \( k \) and distance metric \( D(X_a, X_b) \). For an unlabeled sample \( X_a \), kNN regressor selects \( k \) nearest samples according to
the defined distance metric: $X_1, X_2, ..., X_k$. Suppose their labels are $y_1, y_2, ..., y_k$, then the label of $X_u$ is estimated to be:

$$Y_u = \frac{Y_1 + Y_2 + ... + Y_k}{k}$$  \hspace{1cm} (1)

Suppose there are some noisy labeled samples in $L$, as shown in Figure 2, C is a noisy sample. When only one kNN regressor is employed, suppose an unlabeled sample $X_1$ is labeled then put into $L$. For a sample $X_2$ which is very close to $X_1$, it will suffer from noise more seriously than $X_1$.

![Figure 2. Single regressor affected by noise.](image)

But if two regressors with certain differences are employed and $X_1$ is labeled by another regressor, $X_2$ may suffer from noise only once. So it is wiser to use two regressors to reduce the effect of noise.

For a sample $X_u$ and two kNN regressors $K_1$ and $K_2$, let $\Omega_1 = \{X_{i1}, X_{i2}, ..., X_{ik}\}$ denote the set of $k$-nearest neighboring samples of $X_u$ on $K_1$, and $\Omega_2 = \{X_{21}, X_{22}, ..., X_{2k}\}$ denote the set on $K_2$. If $\Omega_1$ and $\Omega_2$ is not same, $K_1$ and $K_2$ is different on $X_u$. The disagreement level of two regressors can be measured by the difference on labeled set $L$.

Initialize $L_1$ and $L_2$ with $L$, which denote the labeled set of $K_1$ and $K_2$.

**Collaborative Training**

In classification problems, classifiers can provide an estimated probability for every class. Suppose that the probability of a sample $X_1$ belonging to class A is 0.7 and to class B is 0.3, while a sample $X_2$ belonging to class A is 0.9 and to class B is 0.1, then obviously $X_2$ is more confident to be labeled.

Unfortunately, in regression problems, the possible predictions are infinite. The definition of authenticity is the key to this algorithm. Samples with high confidence level should be consistent with the trend of regression, so a sample with high confidence level should make the mean squared error (MSE) of the regressor on the labeled sample set decrease. Since repeatedly measuring the MSE on the whole labeled set is of great computational load, the following method is adopted as an approximation.

For an unlabeled sample $X_a$ in $U$, use $K_1$ to predict its label $Y_a$. Let $\Omega = \{X_{a1}, X_{a2}, ..., X_{ak}\}$ denote the set of $k$-nearest neighboring samples of $X_a$, and their labels are $y_{a1}, y_{a2}, ..., y_{ak}$. Let $K_i(X)$ denote the estimated value of $K_i$ to $X$. Then the error of $K_1$ on $\Omega$ is defined as:

$$E = \sum_{i=1}^{k} (Y_a - K_1(X_a))^2$$  \hspace{1cm} (2)

Let $K_i'$ denote the refined regressor which has utilized the information provided by $(X_a, Y_a)$. The error of $K_i'$ on $\Omega$ is:

$$E' = \sum_{i=1}^{k} (Y_a - K_i'(X_a))^2$$  \hspace{1cm} (3)

Then the confidence level of $X_a$ can be defined as:

$$T_a = E - E'$$  \hspace{1cm} (4)
The sample in $U$ which maximizes $T_a$ is the sample with highest confidence. Let $X_m$ denote this sample. $K_1$ will put $(X_m, Y_m)$ into $L_2$, and in the same iteration $K_2$ will put the sample with the highest confidence into $L_1$.

The following is an analysis of the feasibility of this criterion.

First, assume that $X_m$ is only among the $k$-nearest neighbors of some samples in $\Omega$. In this case, apparently maximizing $T_a$ also makes the MSE of the regressor on the whole labeled set decrease most.

Second, assume that $X_m$ is not among the $k$-nearest neighbors of any samples in $\Omega$. In this case, according to (4) its confidence level is zero, thus it leads to a contradiction.

Third, assume that $X_m$ is among the $k$-nearest neighbors of some samples in $\Omega$ and some other samples not in $\Omega$. In this case it’s hard to evaluate whether $X_m$ can makes the MSE of regressor on the whole labeled set decrease most. Nevertheless, experiments show that in most cases this method is effective.

**Stopping Condition**

Repeat the training process mentioned in section C until any of these conditions occurs:

1) The maximum number of iteration is reached.
2) The maximum amount of time is exceeded.
3) No sample that can make (4) be positive exists in $U$.

Final output for a new sample $X$ is:

$$Y = \frac{K_1(X) + K_2(X)}{2} \quad (5)$$

**Experiments**

The experimental data are from the actual production data of China Oilfield Services Limited in Bohai area.

Two sets of data from different wells are used in the experiment. Each data set is made up of 300 labeled samples and 4700 unlabeled samples. Each sample has 7 attributes: diameter, neutron, acoustic time, Gamma ray, deep resistivity, density, and photoelectric absorption index. All the input attributes have been normalized to $[0.0,1.0]$. The output is porosity.

The k value of regressor $K_1$ is set to 3, and Euclidean distance is used:

$$D(X_a, X_b) = \left( \sum_{i=1}^{d} |X_{a,i} - X_{b,i}|^2 \right)^{1/2} \quad (6)$$

The k value of regressor $K_2$ is also set to 3, but Minkowski distance is used and the p value is set to 5:

$$D(X_a, X_b) = \left( \sum_{i=1}^{d} |X_{a,i} - X_{b,i}|^p \right)^{1/p} \quad (7)$$

The maximum number of iteration $T$ is set to 100, the maximum amount of time is set to 600 seconds. 50% of labeled samples are used for training and the remaining 50% of samples are kept as test set. In each iteration, unlabeled data set $U$ contains 100 samples which are randomly picked.

A nonlinear regression algorithm is tested for comparison, which is usually adopted by manual log interpretation. Moreover, back-propagation neural network is also tested as a representative of supervised learning.

The experiment was repeated five hundred times for each data set. The result is shown in Table 1 and Table 2. Average relative error and correlation coefficient are employed to measure their generalization performance. The data in Table 1 and Table 2 refer to the average value of five hundred runs.
Table 1. Experiment result on set 1.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Average Relative Error</th>
<th>Correlation Coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>Collaborative Regression</td>
<td>12.7%</td>
<td>0.8576</td>
</tr>
<tr>
<td>Nonlinear Regression</td>
<td>20.4%</td>
<td>0.7922</td>
</tr>
<tr>
<td>back-propagation neural network</td>
<td>16.3%</td>
<td>0.8254</td>
</tr>
</tbody>
</table>

Table 2. Experiment result on set 2.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Average Relative Error</th>
<th>Correlation Coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>Collaborative Regression</td>
<td>8.6%</td>
<td>0.9137</td>
</tr>
<tr>
<td>Nonlinear Regression</td>
<td>18.4%</td>
<td>0.8126</td>
</tr>
<tr>
<td>back-propagation neural network</td>
<td>13.2%</td>
<td>0.8679</td>
</tr>
</tbody>
</table>

Table 1 and Table 2 shows that Collaborative Regression has a lower average relative error and a higher correlation coefficient than the other two algorithm, which proves that it can exploit unlabeled samples to improve generalization performance. Secondary interpretation of porosity in log interpretation is a typical application task where labeled samples is insufficient, so the generalization performance of supervised learning methods is likely to be limited. Disagreement-based collaborative regression performs better in such cases.

**Conclusion**

Secondary interpretation of porosity in log interpretation is a typical application task where labeled samples are insufficient. The generalization performance of supervised models is related to the number of labeled samples, so often the accuracy is limited. This paper propose to use Disagreement-based collaborative regression to solve this problem. The algorithm employs two kNN regressors with disagreement, and in every iteration each regressor labels an unlabeled sample with highest confidence level for the other regressor. This algorithm can exploit unlabeled samples to improve regression estimates. The experiment result shows that in both sets disagreement-based collaborative regression is superior to the other algorithms.

This method doesn’t require rich experience and high cost, thus it can avoid the shortage of manual interpretation and be a substitution for it. In the actual production, especially in the newly developed area where labeled data is fairly scarce, it has certain application value.

**References**


