Research on TV imaging casing damage detection and classification method based on C4.5 decision tree

Ying Cuan¹, Zeshang Wang¹*, Jiaxin Han¹

¹School of Computer Science, Xi’an Shiyou Universit, Xi’an Shaanxi, 710065, China
*Corresponding author’s e-mail: ying_cuan@xsyu.edu.cn, zeshawnwang@gmail.com, jxhan@xsyu.edu.cn

Abstract. The casing damage is one of the important problems in oilfield development. The television imaging technology can get enough images of the inner wall of casing to detect the casing, but the efficiency of manually identifying the images is very low. Through the mapping relationship between image and casing damage type, the image features can be obtained. Then, the continuous data threshold selection of C4.5 decision tree algorithm is improved. Finally, the improved C4.5 decision tree algorithm is used to classify casing damage types. Experimental results show that the detection and classification method proposed in this paper can be well implemented.

1. Introduction

Oil is a necessity for industrialization in every country. However, due to time, engineering and geological factors, the casing of oil and gas wells in various oil fields in China have been squeezed, corroded and worn out with high frequency and intensity, leading to serious casing damage and easy deformation, fracture and dislocation. Once the casing problem occurs, it will affect the water injection effect of oil and gas wells, thereby reducing the production, shortening the service life of oil and gas wells, and even causing other quality accidents [1]. However, the efficiency of manual identification of logging data is extremely low, so it is urgent to adopt relevant detection and identification schemes to detect casing and combine them with the classification of damage types, as to determine the subsequent repair or abandonment. Liu ling use 40 arm caliper to get inside the casing 3d model using decision tree algorithm ID3 classifying types of casing damage [2], Zhou Xiangguang and Li Dawei used gradient lifting decision tree to analyze casing environmental parameters, material parameters, production parameters and maintenance parameters and obtain casing damage type [3].Wang Min uses artificial neural network algorithm to analyze the casing address factor and get the casing damage type [4]. Huang ye used BP neural network to analyze the welding seam [5]. Cai Xiaolong analyzed the welding seam based on SVM [6]. However, there is little research on the more intuitive imaging of borehole television. In this paper, downhole television logging technology and C4.5 decision tree algorithm is adopted to overcome the feature transfer between the casing internal section image and the algorithm, improve the node threshold selection in C4.5 algorithm, and realize the recognition of casing damage type.

2. Relevant theories

2.1. Downhole television logging technology

Downhole television logging technology system is composed of two parts: ground instrument and downhole. Under the illumination of the backlit source of downhole instrument, the camera of downhole
TV logging instrument takes pictures of the inner wall of casing and wellbore [7]. After processing, the image signal is transmitted to the ground and displayed on the monitor for operators’ reference. At the same time, it is sent to the video recorder for recording.

Different from ultrasonic tool, television logging technology field practical and display well logging images, without the need to use any operating software or data processing software. It is recorded by video tape and edited by logging images. In the field, images of casing wall and well can be displayed in real time, and necessary engineering schemes can be adopted according to the logging data to shorten the downtime of the construction well [8]

2.2. C4.5 decision tree algorithm

Decision tree is a form of implementing decision rules through tree structure. A decision tree contains one root node, countless leaf nodes and intermediate nodes. The intermediate node corresponds to each decision of attribute value, and the leaf node corresponds to the final decision result. Each parent divides its decision sample set according to the result of attribute decision and assigns it to its children. The root node is the earliest decision place of the total sample set. The path from root to each leaf corresponds to a decision classification rule.

In the learning and establishment of the decision tree, in order to make the best decision effect of each iteration of the decision tree, the optimal partition attribute should be selected from the current sample set for each decision. ID3 algorithm uses the gain of information entropy as the basis for selecting the optimal partition attribute. When the attribute with the largest information gain takes different values, the information amount will increase more and more. Information entropy is a measure of information. Assume that the proportion of the $m$-th sample in the current sample set $S$ is $p_m (m = 1, 2, ..., |S|)$, then the information entropy of is defined as

$$E(S) = -\sum_{m=1}^{V} p_m \log_2 p_m$$  \hspace{1cm} (1)

It is assumed that discrete attribute $a$ has $N$ different values of $\{a^1, a^2, ..., a^N\}$. If $a$ is used to divide the sample data set $S$, $N$ branch nodes will be generated. Among them, the $n$-th branch node contains all samples in $S$ whose values are $a^n$ on attribute $a$, denoted as $S^n$. We can according to the formula (1) calculate the $S^n$ information entropy, again because of the different branch node contains the number of samples is not completely equal, so give each branch node gives the weight of the total sample size $|S^n|/|S|$, namely the sample size of the more than branch node of the greater the influence of the decision-making, then to calculate the property $a$ sample set $S$ division of information gain.

$$G(S, a) = E(S) - \sum_{n=1}^{N} \frac{|S^n|}{|S|} E(S^n)$$  \hspace{1cm} (2)

C4.5 decision tree based on ID3 decision tree is extended, it does not directly using the information gain, but add gain rate as select one of the attributes of the optimal partition standard [9]. Gain rate is defined as

$$R(S, a) = \frac{G(S, a)}{IV(a)}$$  \hspace{1cm} (3)

Where $IV(a)$ is called the "intrinsic value" of attribute $a$ and is defined as

$$IV(a) = -\sum_{n=1}^{N} \frac{|S^n|}{|S|} \log_2 \left( \frac{|S^n|}{|S|} \right)$$  \hspace{1cm} (4)

The larger the number of possible values of attribute $a$ (i.e., the larger $N$ is), the larger the value of $IV(a)$ will normally be[10]. But the use of pure gain rate criterion to preferences can be less number of attribute values, therefore, not directly choose C4.5 algorithm gain rate of the largest differentiate candidate attributes, but combined with information gain and gain rate: first find the attribute of
information gain above average properties, then choose the highest gain rate as the optimal partition properties.

3. Test content

3.1. Downhole television logging technology

In downhole television logging technology, the standard image consists of the casing circular section near the camera, the casing section far away and the casing inner wall between, which maps different types of casing damage. The types of casing damage are classified as follows:

- deformation: the deformation is a sag in the casing.
- dislocation: dislocation means that there is an additional casing truncation surface on the inner wall of the casing in the standard image, and its shape is round or spiral. The inner wall between the fault and the distant casing and the circular cross section of the distant casing together form a larger circle in the television image.
- fracture: the fracture can be divided into two types of casing damage: fracture and perforation. In other words, local damage occurs to the inner wall of casing, which is shown in the standard image as a protruding point at the far edge of casing section.

According to the analysis of casing damage type, the deformation of casing can be identified by the maximum and minimum radii of inner circle and outer circle in TV imaging. The primary data can be obtained by removing noise, binarization and contour processing of the image, and then extracting the maximum and minimum radius of the inner and outer circular contour. Let the maximum radius of the outer circle be \( R_{\text{OMAX}} \), the minimum radius be \( R_{\text{OMIN}} \), and the standard radius be \( R_{\text{OST}} \). The maximum radius of the inner circle is \( R_{\text{IMAX}} \), the minimum radius is \( R_{\text{IMIN}} \), and the standard radius is \( R_{\text{ISTR}} \).

The internal imaging diagram of the casing and its corresponding processing diagram are shown in Fig. 1 and Fig. 2.

\[ \text{Figure 1. Internal imaging diagram of the casing} \]
\[ \text{Figure 2. Processing diagram} \]

On this basis, the following parameters are set to identify the characteristics of casing damage type:

- Radius jump degree \( (R_{\text{IMAX}} - R_{\text{IMIN}}) / R_{\text{IST}} \): this feature can be used to identify fracture. Through this feature, we can judge the jump degree of the maximum radius and minimum radius of the inner circle. When the feature is greater than the threshold, the casing is fractured or deformed.

- Ratio of the maximum radius of the outer circle to the standard radius \( (R_{\text{OMAX}} / R_{\text{OST}}) \): this feature can be used to identify whether the casing is broken near the threshold. If the characteristic value is less than the threshold value, the casing is dislocated.

- Ratio of the minimum radius of the outer circle to the standard radius \( (R_{\text{OMIN}} / R_{\text{OST}}) \): this feature can identify whether the casing is deformed nearby. If the characteristic value is less than the threshold value, the casing is deformed or dislocated.

- Ratio of the maximum radius of the inner circle to the standard radius \( (R_{\text{IMAX}} / R_{\text{ISTR}}) \): this feature can be used to identify whether the casing is broken at a distance. If the eigenvalue is greater than the threshold value, it is fractured or dislocated.

- Ratio of the minimum radius of the inner circle to the standard radius \( (R_{\text{IMIN}} / R_{\text{ISTR}}) \): this feature can identify whether the casing is deformed in the distance. If the characteristic value is less than the
threshold value, the casing is deformed; If the eigenvalue is greater than the threshold value, it is dislocated.

3.2. Downhole television logging technology

In the optimal selection of branch attributes on some code, for continuous attributes, C4.5 algorithm adopts the method of discretization of continuous attributes and arranges $N$ attribute values of attribute $a$ in ascending order. Take the midpoint of the pairwise combination of attribute values in order, and there are $N-1$ segmentation thresholds. Calculate the information gain of these $N-1$ partition methods; The segmentation threshold with the maximum information gain is selected as the judgment threshold of attribute $a$, and the information gain rate in this case is the information gain rate of this attribute.

Although this method has a high accuracy in selecting the optimal segmentation threshold, it requires multiple sequential scanning, calculation and sorting of the data set, resulting in low computational efficiency of the algorithm. As for the characteristic attributes of television imaging, although the values of continuous data of each image feature are different, the values only determine whether the data is in the normal range, i.e., "greater than, normal and less than", so the selection of continuous attribute threshold in C4.5 algorithm can be improved as follows:

- extract the data of "normal" casing type from the training sample data set, and arrange the values of attributes in order to get the maximum value $a_{\text{max}}$ and minimum value $a_{\text{min}}$.
- change the value of each attribute to $(a_{\text{max}}, +\infty)$, $[a_{\text{min}}, a_{\text{max}}]$ and $(-\infty, a_{\text{min}})$ according to the interval "$g$", "$n$" and "$s$".
- repeat the above operation for the remaining properties until the end.

Through the improvement, the attribute segmentation threshold is obtained from the maximum and minimum values of normal data and the continuous data is divided into discrete data. Even though the binary tree model in computer is more efficient than multi-fork tree, the decision tree's conversion from binary tree to tri-tree reduces the efficiency, but the efficiency gains from the improvement of the process efficiency are far greater than the efficiency losses.

4. Results

4.1. Construct decision tree model

The train data set obtained from 30 Wells is shown in Table 1, and the decision tree model is constructed according to the C4.5 decision tree algorithm mentioned above.

According to the sample set, a total of 7 samples were fractured, 6 samples were dislocated, 15 samples were deformed, and 2 samples were normal. After calculation, the total information entropy of the sample set is 1.71, and the information entropy, information gain and information gain rate of each attribute value are shown in Table 2.

Therefore, the attribute with the largest gain "Ratio of the minimum radius of the inner circle to the standard radius" was selected as the root node, and then the decision tree was established from top to bottom.

The path between the root node and leaf node of each decision tree forms a classification rule, and the value of attribute on the path is the rule of decision, and the casing damage type displayed by leaf node is the conclusion of the rule.

| Serial number | Category | Attribute | Attribute | Attribute | Attribute |
|---------------|----------|-----------|-----------|-----------|-----------|
|               |          | Ratio of the maximum radius of the outer circle to the standard radius | Ratio of the minimum radius of the outer circle to the standard radius | Ratio of the maximum radius of the inner circle to the standard radius | Ratio of the minimum radius of the inner circle to the standard radius | Radius jump degree |

Table 1. Train data set
|   | Fracture | 1.125 | 1.113 | 1.053 | 0.929 | 0.125 |
|---|----------|-------|-------|-------|-------|-------|
| 2 | Dislocation | 0.883 | 0.873 | 1.255 | 1.200 | 0.056 |
| 3 | Fracture | 1.222 | 0.891 | 0.916 | 0.663 | 0.253 |
| 4 | Dislocation | 1.043 | 0.964 | 1.192 | 1.133 | 0.059 |
| 5 | Fracture | 1.277 | 1.121 | 1.066 | 0.818 | 0.248 |
| 6 | Fracture | 1.111 | 0.854 | 1.081 | 0.958 | 0.123 |
| 7 | Normal | 1.212 | 1.034 | 1.034 | 0.937 | 0.097 |
| 8 | Dislocation | 1.195 | 0.851 | 1.191 | 1.143 | 0.048 |
| 9 | Normal | 1.120 | 1.090 | 1.007 | 0.962 | 0.045 |
| 10 | Fracture | 1.226 | 0.925 | 1.114 | 0.970 | 0.143 |
| 11 | Deformation | 1.233 | 1.090 | 0.945 | 0.865 | 0.080 |
| 12 | Deformation | 1.218 | 0.889 | 0.932 | 0.789 | 0.143 |
| 13 | Deformation | 1.243 | 1.205 | 0.835 | 0.669 | 0.166 |
| 14 | Dislocation | 1.241 | 1.188 | 1.116 | 1.055 | 0.061 |
| 15 | Deformation | 1.298 | 1.143 | 1.080 | 0.842 | 0.238 |
| 16 | Dislocation | 1.278 | 0.982 | 1.077 | 1.012 | 0.065 |
| 17 | Dislocation | 0.916 | 0.868 | 1.044 | 0.960 | 0.084 |
| 18 | Deformation | 1.121 | 1.083 | 1.038 | 0.775 | 0.264 |
| 19 | Dislocation | 1.288 | 1.079 | 1.086 | 1.019 | 0.067 |
| 20 | Deformation | 1.265 | 1.236 | 1.135 | 0.849 | 0.286 |
| 21 | Deformation | 1.001 | 0.941 | 1.063 | 0.842 | 0.222 |
| 22 | Deformation | 1.073 | 1.000 | 1.005 | 0.921 | 0.083 |
| 23 | Deformation | 1.003 | 0.900 | 1.068 | 0.774 | 0.294 |
| 24 | Fracture | 1.287 | 0.896 | 1.069 | 0.928 | 0.141 |
| 25 | Deformation | 1.027 | 0.913 | 0.928 | 0.647 | 0.281 |
| 26 | Deformation | 1.245 | 1.229 | 1.000 | 0.868 | 0.132 |
| 27 | Deformation | 1.193 | 1.033 | 0.969 | 0.826 | 0.144 |
| 28 | Fracture | 1.265 | 0.928 | 1.072 | 0.992 | 0.080 |
| 29 | Fracture | 1.190 | 0.862 | 1.087 | 0.978 | 0.109 |
| 30 | Deformation | 1.129 | 1.094 | 1.073 | 0.859 | 0.214 |

Table 2. A slightly more complex table within.

| Attribute | Information |
|-----------|-------------|
| Ratio of the maximum radius of the outer circle to the standard radius | 1.63 0.08 38.3% |
| Ratio of the minimum radius of the outer circle to the standard radius | 1.60 0.11 11.3% |
| Ratio of the maximum radius of the inner circle to the standard radius | 1.10 0.61 44.7% |
| Ratio of the minimum radius of the inner circle to the standard radius | 0.96 0.75 80.8% |
| Radius jump degree | 1.36 0.35 37.0% |

4.2. Analysis of casing damage identification results
The result of the decision tree model is analyzed. The selected samples are shown in Table 3.

Table 3. Train data set

| Serial number | Category | Ratio of the maximum radius of the outer circle to the | Ratio of the minimum radius of the outer circle to the | Ratio of the maximum radius of the inner circle to the | Ratio of the minimum radius of the inner circle to the | Radius jump degree | C4.5 decision tree algorithm |
|---------------|----------|-------------------------------------------------------|-------------------------------------------------------|-------------------------------------------------------|-------------------------------------------------------|------------------|--------------------------------|
| 1 Fracture | 1.125 | 1.113 | 1.053 | 0.929 | 0.125 |
| 2 Dislocation | 0.883 | 0.873 | 1.255 | 1.200 | 0.056 |
| 3 Fracture | 1.222 | 0.891 | 0.916 | 0.663 | 0.253 |
| 4 Dislocation | 1.043 | 0.964 | 1.192 | 1.133 | 0.059 |
| 5 Fracture | 1.277 | 1.121 | 1.066 | 0.818 | 0.248 |
| 6 Fracture | 1.111 | 0.854 | 1.081 | 0.958 | 0.123 |
| 7 Normal | 1.212 | 1.034 | 1.034 | 0.937 | 0.097 |
| 8 Dislocation | 1.195 | 0.851 | 1.191 | 1.143 | 0.048 |
| 9 Normal | 1.120 | 1.090 | 1.007 | 0.962 | 0.045 |
| 10 Fracture | 1.226 | 0.925 | 1.114 | 0.970 | 0.143 |
| 11 Deformation | 1.233 | 1.090 | 0.945 | 0.865 | 0.080 |
| 12 Deformation | 1.218 | 0.889 | 0.932 | 0.789 | 0.143 |
| 13 Deformation | 1.243 | 1.205 | 0.835 | 0.669 | 0.166 |
| 14 Dislocation | 1.241 | 1.188 | 1.116 | 1.055 | 0.061 |
| 15 Deformation | 1.298 | 1.143 | 1.080 | 0.842 | 0.238 |
| 16 Dislocation | 1.278 | 0.982 | 1.077 | 1.012 | 0.065 |
| 17 Dislocation | 0.916 | 0.868 | 1.044 | 0.960 | 0.084 |
| 18 Deformation | 1.121 | 1.083 | 1.038 | 0.775 | 0.264 |
| 19 Dislocation | 1.288 | 1.079 | 1.086 | 1.019 | 0.067 |
| 20 Deformation | 1.265 | 1.236 | 1.135 | 0.849 | 0.286 |
| 21 Deformation | 1.001 | 0.941 | 1.063 | 0.842 | 0.222 |
| 22 Deformation | 1.073 | 1.000 | 1.005 | 0.921 | 0.083 |
| 23 Deformation | 1.003 | 0.900 | 1.068 | 0.774 | 0.294 |
| 24 Fracture | 1.287 | 0.896 | 1.069 | 0.928 | 0.141 |
| 25 Deformation | 1.027 | 0.913 | 0.928 | 0.647 | 0.281 |
| 26 Deformation | 1.245 | 1.229 | 1.000 | 0.868 | 0.132 |
| 27 Deformation | 1.193 | 1.033 | 0.969 | 0.826 | 0.144 |
| 28 Fracture | 1.265 | 0.928 | 1.072 | 0.992 | 0.080 |
| 29 Fracture | 1.190 | 0.862 | 1.087 | 0.978 | 0.109 |
| 30 Deformation | 1.129 | 1.094 | 1.073 | 0.859 | 0.214 |
The results of the sample with ordinal number 1 obtained from table 3 are different. After analysis, the error is pruning of the decision tree in order to prevent over-fitting of the decision tree due to too many samples during the construction of the decision tree. In the future research, more representative data should be extracted to make better prediction.

5. Conclusion
Casing damage is one of the important factors affecting oilfield production. This paper aims at the problems such as low efficiency of identifying casing damage in artificial well and easy to be affected subjectively. This paper proposes to combine the downhole television logging technology with the improved C4.5 decision tree algorithm. The type of casing loss is obtained by decision tree algorithm. The experimental results show that the prediction results are roughly in line with the actual situation, which shows the rationality of the model construction and achieves the purpose of improving efficiency.

Acknowledgments
This work was supported by the Key Research Development Program of Shaanxi Province (2019KW-045) and the National Natural Science Foundation of China (51707158).

References
[1] Chen Lu. Discussion on the new technique of casing damage detection, Management & Technology of SME, 2016(02):286.
[2] Liu ling. The study and realization of the system of the casing 3D visualization and casing damage identification, Northeast Petroleum University,2014.
[3] Zhou Xiangguang and Li Dawei. Prediction of casing failure by gradient boosting decision tree algorithm, Journal of Computer Applications,2018,38(S2):144-147.
[4] Wang Min. Application research of neural network method in casing damage prediction, Northeast Petroleum University,2007.
[5] Huang Ye. Welding defects modeling and recognition algorithm reasearching based on Back-Propagation neural network, Xi’an Shiyou University,2016.
[6] Cai Xiaolong. Welding defects modeling and recognition algorithm reasearching based on Support Vector Machine, Xi’an Shiyou University,2014.
[7] Chai Manzhuo, Xiang Xujin, Zhang qingsheng and Yang Qingrong. Applications of Downhole Video Logging Tool in Casing Inspections, Well Logging Technology,2002(03):242-246+264.
[8] Wang Yunqiang, Yang Qingxian, Wang Lei and Suo Fei. The failure of oil and water well is detected by eagle - eye TV combination logging technique, Inner Mongolia Petrochemical Industry,2003(S1):138-139.
[9] S R Istiana and I Waspada. Using C4.5 algorithm to predict students monthly payment on islamic boarding school, Journal of Physics: Conference Series,2019,1217(1).
[10] Farisa Hafida Syahrial, Budhi Irawan and Anggunmeka Luhur Prasasti. Detecting Hand, Foot and Mouth Disease in Earlier Stage Using C4.5 Algorithm as Expert System Based on Android, Journal of Physics: Conference Series,2019,1201(1).