

A PREDICTION APPROACH TO WELL LOGGING

Qing He¹, Ping Luo¹, Zhong-Zhi Shi¹, Yalei Hao² and Markus Stumptner²

1. The Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing, 100080)

2. Advanced Computing Research Centre, University of South Australia, SA5095, Australia

Abstract: How to provide a means or organize the information used in making exploration decisions in petroleum exploration is an important task. In this paper, a machine learning method is put forward to collect experiences and estimate or prediction the absent data. The well logging experiments show that the method is efficiently and accurately.

Key words: hyper surface, petroleum exploration, well logging

1. INTRODUCTION

According to the Encyclopedia Britannica, around 547 billion barrels of petroleum remain undiscovered on Earth (plus known reserves of 1,119 billion barrels). Experts estimate we will run out of petroleum in 50 years at the current rate of consumption, around 60 million barrels per day. These facts drive the highly competitive oil companies to make their exploration more efficient.

How to provide a means or organize information used in making exploration decisions in petroleum exploration is an important task. An explorationist, who must recommend the wildcat well location and the payzone for the well, is making a decision which might cost a company

millions of dollars. Thus, one who is able to assimilate and use the greatest amount of data is making a more informed decision.

Data about a potential petroleum-bearing area, i.e., a payzone, are obtained from various quantitative sources such as seismic instruments and well loggings and outcrop data, field notes, and maps. Large computer programs for analyzing numeric data already exist. The petroleum industry is today the largest industrial user of supercomputers. These programs provide the best quantitative data about the payzone. Those results, together with all of the other data collected about a payzone fill corporate database. Unfortunately, the amount of data is so great and diverse that it is not always accessible for the purpose of the payzone analysis in a reasonable amount of time. Before making a recommendation concerning the drilling target for oil or gas, the explorationist must "interpret" the data, i.e., making an analysis of its potential to produce oil or gas. At this point the experienced petroleum explorationist is at an advantage because he has interpreted other payzones. He draws very strongly upon his knowledge of other payzones, drawing as many as possible from his background. Moreover, these data has two inherent problems due to the amount and the nature of data available to the explorationist, leading to the difficulty in attempting to interpret a payzone. Firstly, data sets can easily reach 10 to 20 gigabytes and larger. Secondly, until recently, no one had a standard and comprehensive way to see the information. Therefore, no one could exploit the data's full value. For example, 3D seismic data reservoirs are characterized by 20 to 30 variables. It is very difficult to comprehend the data. For some payzones the corporate database overflow with data, and in others very little data are available. Sometimes data conflict and the explorationist must decide which data he believes to be of higher quality. Decisions are often made on the basis of incomplete and inaccurate data of dubious quality.

Classification method can be used for analysis well logging data. There are many classification methods have been studied in the past. Vapnik and his research group have studied machine learning based on finite samples since sixtieths in the last century. A complete theory, Statistical Learning Theory, has been established (see ref. 1-5). Moreover, a new universal learning algorithm SVM, Support Vector Machine, has been developed. It is an especially efficient classification algorithm for finite samples, nonlinear and high dimension data. The main idea of SVM is mapping the nonlinear data to a higher dimension linear space where the data can be linear classified by hyper plane. The mapping is a nonlinear mapping defined by inner product function. So a lot of repeat computations of inner product for m -matrices cannot be avoided, where m is the number of samples. It is almost impossible to classify large data more than 4000 samples by using PC (see ref. 6).

In fact, SVM maps the data into higher dimension linear space in which the data can be separated by a hyper plane. It is show that SVM is an indirectly method to solving nonlinear problem.

It is an important problem how to find a classification method, which has neither increasing dimension mapping nor computing a large matrix product. In paper (see ref. 7), a novel classification method based on hypersurface is put forward to solve the problem. Jordan Curve Theorem is the theoretic base of classification method based on hyper surface.

The next sections are arranged as following:

Section 2 introduces HSC i.e. classification method based on hypersurface, and put forward a prediction method. Section 3 put forward a method to display the petroleum data based hypersurface. Section 4 put forward a data prediction method and give some results of well logging experiments. Finally, section 4 contains the conclusions.

2. THE PREDICTION METHOD

The Jordan Curve Theorem. Let X be a closed set in R^3 . If X is homeomorphism to a sphere S^2 , then its complement $R^3 \setminus X$ has two connected components, one bounded, the other unbounded. Any neighborhood of any point on X meets both of these components (see ref. 8).

The Jordan Curve Theorem shows that any 2D closed surface obtained from sphere by conscious transform separates three dimension space into two regions, one called inside, and the other is called outside. The surface can be used to classify data. This kind of surface is called hyper surface. For any given point, how to determine it is inside or outside about the hyper surface is the first important problem.

Classification Theorem. Let X be a closed set in R^3 . If X is a homeomorphism to a sphere S^2 , then its complement $R^3 \setminus X$ has two connected components, one called inside, the other called outside. For any $x \in R^3 \setminus X$, then the point x is inside of $X \Leftrightarrow$ the intersecting number between any radial from x and X is odd; x is outside of $X \Leftrightarrow$ the intersecting number between any radial from x and X is even.

Theorem (Jordan Curve Theorem in High Dimension Space). Suppose that $X \subset S^n$ is homeomorphism to a sphere S^m , then $m \leq n$, otherwise $X = S^n$. If $m < n$, then the homology group of $S^n \setminus X$ is

$$H_k(S^n \setminus X) \cong \begin{cases} Z \oplus Z, & \text{if } m = n - 1 \text{ and } k = 0, \\ Z, & \text{if } m < n - 1 \text{ and } k = 0, \\ 0, & \text{otherwise.} \end{cases}$$

Specially, if $m = n - 1$, then $S^n \setminus X$ composed by two connected components, and if $m < n - 1$, then there exists only one connected component.

Based on the Jordan Curve Theorem, space can be separated by a two-sided surface that is a homeomorphism to a sphere. The separating hyper surface may be consisting of many hyper surfaces. For any given point, how to decide whether it is inside of the separating or outside depend on that the intersecting number between the separating hyper surface and the radial from the point is odd or even. This classification method is a direct and convenient method. But how to construct the separating hyper surface is an important problem. Based on the Jordan Curve Theorem, we put forward HSC, i.e. classification method based on separating hyper surfaces in ref. 7.

Step1. Find a rectangle region that includes all samples.

Step2. Divide the region into some smaller regions, which only contain at most one sample.

Step3. Label each region according the inside sample's class. Then the frontier vectors and the class vector form a link.

Step4. Combine the same class connected region and obtain a separating hyper surface then saved it as a link.

Step5. Input a new sample and calculate the intersecting number of separating hyper surfaces for the sample. Draw a radial from the sample. Then according to whether the intersecting number between the radial and the separating hyper surface is even or odd, the class of the sample is decided.

Note: If there are different training samples in the same unit then divide the lattice again by unit lattices and do the same things as step 1 until all samples have been used. So a hyper surface formed and denoted by a string. This is called local elaborate strategy.

Suppose that there are d condition attributes and m values about some decision attribute given by the explorationist or obtained from experiment data such as well logging. To learn and collect the explorationist's experiences included in the data, we think that the data samples belong to m classes. The following training algorithm is put forward.

Step1. Input m classes' training samples. Let the training samples be distributed in the rectangle region.

Step2. Transform the region into unit region.

Step3. Divide the region into 10^d equal small regions.

Step4. Label the small regions by 1, 2..... m according to the inside samples' class.

Step5. Remerge the frontiers of the same class regions in where is the same class then save it as a link table.

Step6. For any regions where there are training samples from different classes in the same unit, go to Step2.

Step7. Repeat the above until there is no region where there are different classes.

At the end, some separating hyper surfaces, which are described by the link tables, are obtained.

After learning from the training sample, the explorationist's classification experience is collected in the hyper surface, i.e. link table. Using hyper surface the absent data can be estimated or predicted from data such as well logging. The steps are as following.

Step1. Input a testing sample and make a radial from the sample.

Step2. Input all link tables of class k ($k = 1,2,3,\dots,m$) obtained by the above training algorithm.

Step3. Count the intersecting number of the sample with the above link table.

Step4. If the intersecting number of the sample with the above link tables is odd then label the sample by k . It is mean that the prediction value is the k th decision value, otherwise go to next step.

Step5. Input all link tables of class $k + 1$ obtained by the above training algorithm. Do step3-4 until $k = m$.

Step6. Calculate the classifying accuracy rate.

This is a universal prediction method for large nonlinear data bases. In fact, For large data sets (10^7) (see Table 1 and Table 2), the speed of HSC is very fast. The reason is that the time of saving and extracting hyper surfaces is very short and the need for storage is very little, which is not the advantage of SVM. Another reason is that the decision process is very easy by using the Jordan Curve Theorem.

Table 1. Training results

Training Samples	Training Time	Recall Time	The Rate of Recall (%)
3,314	1s	2s	100.00
6,677	3s	4s	100.00
9,525	4s	7s	100.00
9,530	6s	7s	100.00
17,919	8s	12s	100.00
43,217	18s	31s	100.00
106,344	48s	1m 17s	100.00
1,053,125	9m 0s	12m 29s	100.00

Table 2. Testing results

Training Samples	Testing Samples	Testing Time	Precision of Prediction (%)
4,735	17,919	12s	99.29
9530	85189	1m	99.89
9,525	17,919	13s	100.00
9,525	43,217	30s	100.00
9,525	106,344	1m 14s	100.00
9,525	1,053,125	12m 19s	100.00

The experiment results in Table 2 of training on small scale samples and testing on large scale show that HSC has strong ability of generalization. If the scale of unit is too small then the hyper space may be separated into two parts. In a region with scattered samples, big unit is required, on the contrast at the region with highly sample density small unit is required. But the samples are not always uniformly distributed. So the locally decided elaborate division is an important strategy. The local division strategy improves the generalization ability and accuracy.

3. DISPLAY AND PREDICT THE DISTRIBUTION OF OIL/GAS

After explorationists interpret the data from wildcat wells in a certain field, they may point out which well may be drilled out oil or gas and which well may not be drilled oil or gas. The critical problem is how to organize the decision information and apply the decision information to predict other point where no well has been drilled? HSC can help us do this.

HSC can solve the nonlinear classification problem which the samples distribute in any shape in a finite region. The method has nothing to do with the distribution of the data, even though the shape is interlock or crisscross. This advantage makes it useful for petroleum data. Figure 1 shows a slice of oil / gas distribution simulation data for a certain formation. Let isotropic homogeneity of the oil / gas field, it was created by our algorithm based on the wildcat data and regional petroleum geologic background. The domain marked blue represents domain where oil or gas may be drilled out. On the contrary, the domain marked red represents domain where no petroleum may be drilled out. HSC can obtain two hypersurface as shown in Figure 2. The hypersurface can be used to predict whether or not oil or gas exists at certain other points where no well exists.

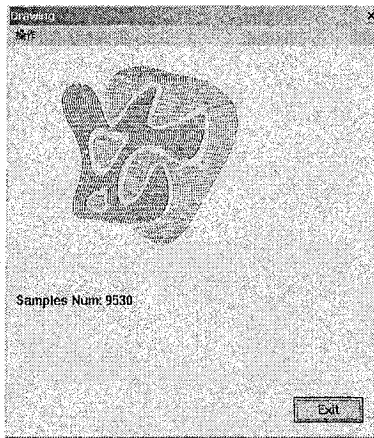


Figure 1. A Slice of Distribution of Petroleum

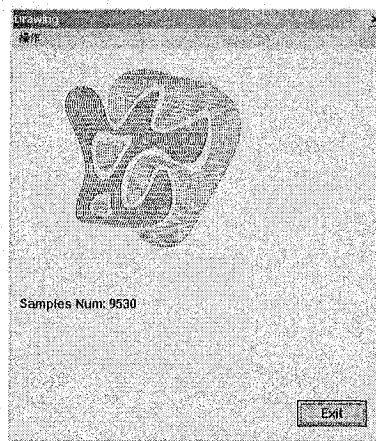


Figure 2. Hypersurface

4. PREDICTION OF ABSENT DATA

Consider the sample data from Figure 3. They are described by $(x, y, z) = (\text{DEPTH}, \text{GR}, \text{AC})$ and DEN which are four kinds of data from well logging. DEPTH is the well depth, GR is natural gamma, AC is acoustic wave time difference, and DEN is density. Now there are 4903 records or data samples about the above four attributes from one prospect, as shown in Figure 4. But only 4408 records are complete, the others lack DEN

data. DEN is an important value in petroleum exploitation. DEN data range from 1.0g/cm^3 to 3.5g/cm^3 . How to predict the absent DEN values is a problem. The problem can be solved by using HSC. The first, we divide the DEN values into 30 classes. The Second, training the above complete data by using the training algorithm. Obtaining the hypersurface as shown in Figure 4 the absent data's class can be predicted. So the DEN value corresponding to the class is decided. In this test correct rate is 94.80%. Figure 3 shows the distribution of samples.

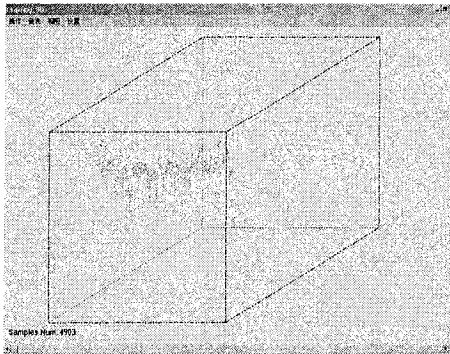


Figure 3. The Distribution of Samples

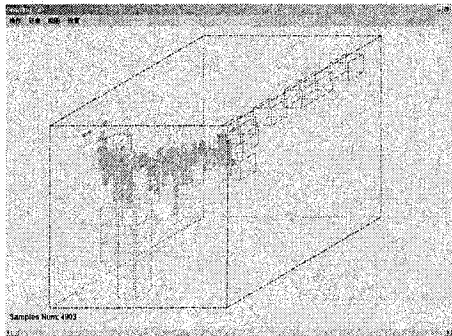


Figure 4. Prediction Hypersurface

5. CONCLUSIONS

In this paper, a prediction method HSC is put forward to collect explorationist's experiences and estimate or prediction the absent data. The

well logging experiments show that the new method is efficiently and accurately. The method has the following advantages.

- 1) High efficient and accuracy;
- 2) Strong ability of prediction;
- 3) Robustness, i.e. data noise can be controlled in a local region;
- 4) Independent on sample feature;
- 5) Multi-values prediction.

Because of these properties, HSC is useful for predicting the distribution of petroleum and absent data.

ACKNOWLEDGEMENTS

This work is supported by the National Science Foundation of China (No.60173017, 90104021), the 863 Project (No.2003AA115220) and China-Australia Special Fund for S&T Cooperation.

REFERENCES

1. Vapnik V. N. Support Vector Method for Function Approximation, Regression Estimation and Signal Processing. Neural Information Processing Systems, Vol.9. MIT Press, Cambridge, MA.
2. Vapnik V. N. The Nature of Statistical Learning Theory. New York: Springer-Verlag, 1995.
3. Vapnik V N.,Levin E,Le Cun Y.Mearsuring the VC-Dimension of learning machine. Neural Computation,1994,6:851-876.
4. Vapnik V. N. Statistical Learning Theory, J. Wiley, New York, 1998.
5. Burges C J C. A tutorial on support vector machines for pattern recognition. Data Mining and Knowledge Discovery, 1998, 2(2).
6. Schokopf,B.,Burges,J.C.,Smola,A.J. Advances in Kernel Methods Support Vector Learning. Cambridge, MA: MIT Press, 1999.
7. Qing He, Zhong-Zhi Shi, Li-An Ren, E. S. Lee. A Novel Classification Method Based on Hyper Surface. International Journal of Mathematical and Computer Modeling,38(2003),395-407
8. William Fulton, Algebraic Topology A First Course, Springer-Verlag 1995.