# INTELLIGENT TECHNOLOGY FOR WELL LOGGING ANALYSIS

Zhongzhi Shi1 Ping Luo1 Yalei Hao2 Guohe Li3

Markus Stumptner2 Qing He1 Gerald Quirchmayr2 4

*1 Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100080, China*

*shizz@ics.ict.ac.cn, luop@ics.ict.ac.cn, heq@ics.ict.ac.cn*

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

*Yalei.Hao@postgrads.unisa.edu.au, mst@cs.unisa.edu.au, Gerald.Quirchmayr@unisa.edu.au*

*3University of Petroleum,Beijing 100080,China*

*guoheli@ sina.com.cn*

*4 Institut für Informatik und Wirtschaftsinformatik, Universität Wien, Liebiggasse 4, A-1010 Wien, Austria*

Abstract: Well logging analysis plays an essential role in petroleum exploration and exploitation. It is used to identify the pay zones of gas or oil in the reservoir formations. This paper applies intelligent technology for well logging analysis, particular combining data mining and expert system together, and proposes an intelligent system for well log analysis called IntWeL Analyzer in terms of data mining platform MSMiner and expert system tool OKPS. The architecture of IntWeL Analyzer and data mining algorithms, including Ripper algorithm and MOUCLAS algorithm are also presented. MOUCLAS is based on the concept of the fuzzy set membership function that gives the new approach a solid mathematical foundation and compact mathematical description of classifiers. The aim of the study is the use of intelligent technology to interpret the pay zones from well logging data for the purpose of

reservoir characterization. This approach is better than conventional techniques for well logging interpretation that cannot discover the correct relation between the well logging data and the underlying property of interest.

Key words: Intelligent Technology, Well Log Analysis, Data Mining, MOUCLAS Algorithm

# 1. INTRODUCTION

In the oil/gas exploration and exploitation well logging plays an essential role. Well logs record measurements of the rock's physical properties as well as the properties of the fluids contained in the rock. So-called well logging is a method of solving geological mission by means of physical principle. Since well logging detects the formation of a borehole from surface down to bottom, well logging data records all the formations' information about lithology, porosity, permeability, oil saturation and so on. According to the measurements of well logging there are three different kinds of data: electrical, nuclear and acoustic. Electrical logging is used to analyze oil saturation and water saturation of the formation. Nuclear logging is used to analyze the porosity and permeability. Acoustic logging is used to determine the amount of clay grain size.

Well logging interpretation includes qualitative and quantitative interpretation. Qualitative interpretation is done by interpreters traditionally. For example, when logging interpreters find the acoustic curve extent great which means porosity great, the neutron curve extent great which means hydrogen great, the gamma ray curve great which means amount of clay lower, they can determine this formation is a pay zone. After the qualitative interpretation, logging interpreters locate the pay zones and calculate the water saturation, oil saturation and amount of clay more accurately.

In order to make the well logging interpretation more efficient and automatic, intelligent technology is popularly adopted. The intelligent technique is fully using the nonlinear transformation to process information, and map the well logging data vectors to the formation features of lithology, porosity, permeability, oil saturation and so on.

Data mining based classification aims to build accurate and efficient classifiers not only on small data sets but more importantly also on large and high dimensional data sets, while the widely used traditional statistical data analysis techniques are not sufficiently powerful for this task. With the development of new data mining techniques on association rules, new classification approaches based on concepts from association rule mining are emerging. These include such classifiers as ARCS[1], CBA[2], LB[3], CAEP[4], etc.,

which are different from the classic decision tree based classifier C4.5[5] and k-nearest neighbor[6] in both the learning and testing phases. In this paper first we investigate inductive learning algorithm Ripper for well logging data analysis. Then according to the characteristic of well logging data we present a new approach to the classification over quantitative data in high dimensional databases, called MOUCLAS (MOUntain function based CLASsification), based on the concept of the fuzzy set membership function. It aims at integrating the advantages of classification, clustering and association rules mining to identify interesting patterns in selected sample data sets.

## 2. BASIC WELL LOG CURVES

In this section we introduce some traditional well log curves and its physical significance.

**Resistivity curve**: Salt water will conduct electricity. However, Oil is an insulator and will not conduct electricity. An electrical curve measures the change in conductivity when tool is pulled up the well bore. Resistivity is the reciprocal of conductivity.

**Gamma ray curve**: The Gamma Ray device measures naturally occurring radiation given off by the rocks. Shale formations have more organic material so they emit more radiation. But sandstones, limestones, and salts have low organic materials. Thus, they have low Gamma Ray counts.

**Neutron curve**: The neutron log is radiation detector that measures the presence of hydrogen atoms in the formation. This indicates pore space by the assumption that the pores will be filled with water or hydrocarbons. The Neutron tool is more sensitive to statistical fluctuations and tends to be more "nervous" than other curves.

**Acoustic curves**: The sonic tool was the first tool developed to measure porosity. It consists of a transmitter that sends out a sound wave into the formation and a receiver, which detects that sound after a certain amount of time has passed. The transmitter and the receiver get out of synchronization with each other. This effect is called cycle skipping. Cycle skipping is a series of too early and too late arrivals back to the receiver, this causes a high to low zig-zag effect on the log. It is a sure indicator of gas in sandstones.

All the curves mentioned above and some other curves are the source of our conditional attributes data in the following experiments using intelligent data mining techniques.

## 3. ARCHITECTURE OF INTWEL ANALYZER

It is useful to apply intelligent technology for well log data analysis. Intelligent well log analysis system called IntWeL Analyzer, which is shown in Figure 1, has been developed. IntWeL Analyzer consists of eight modules, including data warehouse, on line analysis processing (OLAP), data mining, knowledge acquisition, visualization, inference engine, knowledge base and interface.

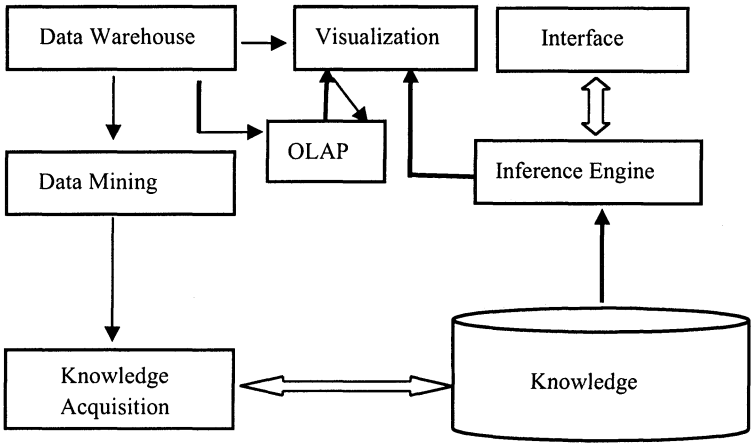


*Figure 1. Architecture of IntWeL Analyzer*

In IntWeL Analyzer first well log data is transferred into data warehouse, then it can be handled by on line analysis processing (OLAP module). Under data warehouse supporting, data mining module will extract knowledge from log data[8]. Knowledge acquisition module will convert data mining results into knowledge base. Inference engine module deploys knowledge to analyze well log data supported by knowledge base[9]. Users can interactive with the IntWeL Analyzer through interface.

We propose the object-oriented knowledge representation for the system. The knowledge will be represented in frame and semantic network based on object-oriented technology. This approach has all the features of an object-oriented mechanism, such as encapsulation, inheritance and message processing. The system puts all production rules into method slots.

In order to provide a powerful inference mechanism, as well as the maximum flexibility and convenience, the system proposes a high-level language, that is Inference Control Language (ICL), which can be used to describe knowledge and rules, and control the inference process.

## 4. INDUCTIVE LEARNING

Inductive learning is an attractive research area in machine learning currently. The set of production rules generated by inductive learning is understandable to humans and this is one important reason we use it for the purpose of reservoir characterization. Among various inductive learning algorithms Ripper (Repeated Incremental Pruning to Producing Error Reduction) is one of the most effective and efficient. In this section we introduce Ripper algorithm first and then use it to analyze the well logging data in order to identify the pay zones of oil.

Ripper algorithm proposed by Cohen in 1995[10]. The underpinning of Cohen's algorithms is a descendant of REP (Reduced Error Pruning) which is a technique used in conjunction with a rule learning system in order to improve the accuracy of a generated rule set[11]. The whole algorithm of Ripper is consisted of two phases: the first is to determine the initial rule set and the second is post-process rule optimization.

1) Generating the initial rule set

This sub-algorithm is described below:

```
procedure Rule_Generating(Pos,Neg)
begin
    Ruleset := {}
    while Pos . {} do
        /* grow and prune a new rule */
        split (Pos,Neg) into (GrowPos,GrowNeg) and
    (PrunePos,PruneNeg)
        Rule := GrowRule(GrowPos,GrowNeg)
        Rule := PruneRule(Rule,PrunePos,PruneNeg)
        if the terminal conditions satisfy then
            return Ruleset
        else
            add Rule to Ruleset
            remove examples covered by Rule from (Pos,Neg)
        endif
    endwhile
    return Ruleset
end
```

The above is a separate-and-conquer rule-learning algorithm. First the training data are divided into a growing set and a pruning set. Then this

algorithm generates a rule set in a greedy fashion, a rule at a time. While generating a rule Ripper searches the most valuable rule for the current growing set in rule space that can be defined in the form of BNF. Immediately after a rule is extracted on growing set, it is pruned on pruning set. After pruning, the corresponding examples covered by that rule in the training set (growing and pruning sets) are deleted. The remaining training data are re-partitioned after each rule is learned in order to help stabilize any problems caused by a "bad-split". This process is repeated until the terminal conditions satisfy.

After each rule is added into the rule set, the total description length, an integer value, of the rule set is computed. The description length gives a measure of the complexity and accuracy of a rule set. The terminal conditions satisfy when there are no positive examples left or the description length of the current rule set is more than the user-specified threshold.

2) Post-process rule optimization

Ripper uses some post-pruning techniques to optimize the rule set. This optimization is processed on the possible remaining positive examples. Re-optimizing the resultant rule set is called RIPPER2, and the general case of re-optimizing "k" times is called RIPPER*k*.

It is clear that the algorithm above is for the binary class attribute problem. Based on this Cohen used a technique called sequential covering to solve the multiple class attribute problem. After arranging the sequence of classes Ripper finds rules to separate $\text{Class}_1$ from $\text{Class}_2,\cdots,\text{Class}_n$, then rules to separate $\text{Class}_2$ from $\text{Class}_3,\cdots,\text{Class}_n$, and so on. The final class $\text{Class}_n$ will become the default. The sequence of the classes can be fixed by one of the followings: increasing frequency of class, decreasing frequency of class, optimal order using heuristic or the user-specified order. Consequently, the end result is that rules for a single class will always be grouped together. When predicting a sample if an example is covered by rules from two or more classes, then this conflict is resolved in favor of the class that comes first in the ordering.

Simplified oil/gas formation identification is a typical classification problem in data mining. It is used to identify the pay zones of gas or oil in the reservoir formation and helps to optimize production of oil/gas reservoir.

We analyze the data in petroleum database that contains three kinds of well logging quantitative data, including electrical, nuclear and acoustic data. The data in our application is from the boreholes in Xinjiang Province, China. There are 4 different layers under the ground: water, mixture of water and oil, oil and shale. They are the class labels of the record at certain depth.

Altogether, every 0.125 meters we get a record with 9 condition attributes and 1 class label. No depth information is used for training.

It is very important to mention that in oil/gas characterization problem the model extracted from the data in one area is not fit for the other area. Thus, we can't use the model from one area to predict the samples in another area. In addition, the training subset should cover the horizontal and vertical boreholes of producing wells. Thus, we combine all the records of different boreholes in one area into one data set and divide it into two parts: one is for training and the other is for testing. We divide the original whole data randomly in order to maintain the same proportion of different classes. The parameters of Ripper for training and the testing results are shown Table 1 and Table 2 respectively.

*Table 1.* Parameters description

| No. | Parameter description |
|---|---|
| 1 | arranging the sequence of classes in increasing frequency of class |
| 2 | using the default rule searching space |
| 3 | each rule to cover at least 10 samples |

*Table 2.* Testing results

| No. | Proportion of records in training set | Proportion of records in testing set | Accuracy |
|---|---|---|---|
| 1 | 2/3 | 1/3 | 93.10% |
| 2 | 1/3 | 2/3 | 93.53% |

Table 2 shows the testing accuracy is satisfactory. And we argue that Ripper really fits for the oil/gas identification application in well logging analysis.

## 5. MOUCLAS ALGORITHM

The *MOUCLAS* algorithm, similar to ARCS, assumes that the initial association rules can be agglomerated into clustering regions, while obeying the anti-monotone rule constraint. Our proposed framework assumes that the training dataset $D$ is a normal relational set, where transaction $d \in D$. Each transaction $d$ is described by attributes $A_j$, $j = 1$ to $l$. The dimension of $D$ is $l$, the number of attributes used in $D$. This allows us to describe a database in terms of volume and dimension. $D$ can be classified into a set of known classes $Y$, $y \in Y$. The value of an attribute must be quantitative. In this work, we treat all the attributes uniformly. We can treat a transaction as a set of (attributes, value) pairs and a class label. We call each (attribute, value) pair an item. A set of items is simply called an itemset.

Since CBA indicates the feasibility of setting up a link between

association rule and classification and ARCS proves that the idea of designing a classification pattern based on clustering can work effectively and efficiently, we design a *MOUCLAS* Pattern (so called *MP*) as an implication of the form:

$$Cluster(D)_t \rightarrow y,$$

where $Cluster(D)_t$ *is a cluster of D,* $t = 1$ to $m$, and $y$ is a class label. The definitions of *frequency* and *accuracy* of *MOUCLAS* Patterns are defined as following: The *MP* satisfying minimum support is ***frequent***, where *MP* has support s if s% of the transactions in *D* belong to $Cluster(D)_t$ and are labeled with class $y$. The *MP* that satisfies a pre-specified minimum confidence is called ***accurate***, where *MP* has confidence c if c% of the transactions belonging to $Cluster(D)_t$ are labeled with class $y$.

Though framework of support – confidence is used in most of the applications of association rule mining, it may be misleading by identifying a rule $A \Rightarrow B$ as interesting, even though the occurrence of A may not imply the occurrence of B. This requires a complementary framework for finding interesting relations. Correlation is one of the most efficient interestingness measures other than support and confidence. Here we adopt the concept of reliability to describe the correlation. The measure of reliability of the association rule $A \Rightarrow B$ can be defined as:

$$\text{reliability} \quad R(A \Rightarrow B) = \left| \frac{P(A \wedge B)}{P(A)} - P(B) \right|$$

Since R is the difference between the conditional probability of B given A and the unconditional of B, it measures the effect of available information of A on the probability of the association rule. Correspondingly, the greater R is, the stronger *MOUCLAS* patterns are, which means the occurrence of $Cluster(D)_t$ more strongly implies the occurrence of $y$. Therefore, we can utilize reliability to further prune the selected *frequent and accurate and reliable MOUCLAS* patterns (*MPs*) to identify the truly interesting *MPs* and make the discovered *MPs* more understandable. The *MP* satisfying minimum reliability is ***reliable***, where *MP* has reliability defined by the above formula.

Given a set of transactions, *D*, the problems of *MOUCLAS* are to discover *MPs* that have support and confidence greater than the user-specified minimum support threshold (called *minsup*), and minimum confidence threshold (called *minconf*) and minimum reliability threshold (called *minR*) respectively, and to construct a classifier based upon *MPs*.

The classification technique, *MOUCLAS*, consists of two steps:

1. Discovery of *frequent*, *accurate* and *reliable MPs*.
2. Construction of a classifier, called *De-MP*, based on *MPs*.

The core of the first step in the *MOUCLAS* algorithm is to find all *cluster_rules* that have support above *minsup*. Let *C* denote the dataset *D*

after dimensionality reduction processing. A *cluster_rule* represents a *MP*, namely a rule:

$$cluset \rightarrow y,$$

where *cluset* is a set of itemsets from a cluster $Cluster(C)_t$, $y$ is a class label, $y \in Y$. The support count of the *cluset* (called *clusupCount*) is the number of transactions in $C$ that belong to the *cluset*. The support count of the *cluster_rule* (called *cisupCount*) is the number of transactions in $D$ that belong to the *cluset* and are labeled with class $y$. The *confidence* of a *cluster_rule* is (*cisupCount* / *clusupCount*) × 100%. The support count of the *class y* (called *clasupCount*) is the number of transactions in $C$ that belong to the class $y$. The *support* of a *class* (called *clasup*) is (*clasupCount* / $|C|$) × 100%, where $|C|$ is the size of the dataset $C$.

Given a *MP*, the *reliability* R can be defined as:

$$R(cluset \rightarrow y) = \left| (cisupCount / clusupCount) - (clasupCount / |C|) \right| \times 100\%$$

The traditional association rule mining only uses a single *minsup* in rule generation, which is inadequate for many practical datasets with uneven class frequency distributions. As a result, it may happen that the rules found for infrequent classes are insufficient and too many may be found for frequent classes, inducing useless or over-fitting rules, if the single *minsup* value is too high or too low. To overcome this drawback, we apply the theory of mining with multiple minimum supports in the step of discovering the frequent MPs as following.

Suppose the total support is *t-minsup*, the different minimum class support for each class $y$, denoted as $minsup_i$ can be defined by the formula:

$$minsup_i = t\text{-}minsup \times \text{freqDistr}(y)$$

where, freqDistr($y$) is the function of class distributions. *Cluster_rules* that satisfy $minsup_i$ are called *frequent cluster_rules*, while the rest are called *infrequent cluster_rules*. If the *confidence* is greater than *minconf*, we say the *MP* is *accurate*.

The task of the second step in *MOUCLAS* algorithm is to use a heuristic method to generate a classifier, named *De-MP*, where the discovered *MPs* can cover $D$ and are organized according to a decreasing precedence based on their confidence and support. Suppose $R$ be the set of *frequent*, *accurate* and *reliable MPs* which are generated in the past step, and $MP_{default\_class}$ *denotes* the default class, which has the lowest precedence. We can then present the *De-MP* classifier in the form of

$$<MP_1, MP_2, \ldots, MP_n, MP_{default\_class}>,$$

where $MP_i \in R$, $i = 1$ to $n$, $MP_a \succ MP_b$ if $n \geq b > a \geq 1$ *and* $a, b \in i$, $C \subseteq \cup$ *cluset of* $MP_i$,.

## 6. CONCLUSIONS

A novel architecture of IntWeL Analyzer for well logging data analysis is proposed in the paper. The IntWeL Analyzer integrates data mining and expert system together. An inductive learning algorithm called Ripper and a novel association rule classification algorithm called MOUCLAS are investigated in this paper for intelligent well logging data analysis.

For the future we will add more expert experiences and geophysical knowledge into the IntWeL Analyzer and attempt to establish a relationship between different well logs, such as seismic attributes, laboratory measurements and other reservoir properties.

## 7. ACKNOWLEDGEMENT

This work was partially supported by the Australia-China Special Fund for Scientific and Technological Cooperation under grant CH030086, the joint China-Australia project under the bilateral scientific exchange agreement between The Chinese Academy of Sciences and Australia Academy of Science.

## REFERENCES

1. B. Lent, A. Swami, and J. Widom. Clustering association rules. ICDE'97, (1997) 220-231
2. B. Liu, W.Hsu, and Y.Ma. Integrating classification and association rule mining. KDD'98. (1998) 80-86
3. Meretakis, D., & Wuthrich, B. Extending naive Bayes classifiers using long itemsets. Proc. of the Fifth ACM SIGKDD. ACM Press. (1999) 165-174
4. Dong, G., & Li, J. Efficient mining of emerging patterns: Discovering trends and differences. Proc. of the Fifth ACM SIGKDD. (1999)
5. Quinlan, J. R. C4.5: Programs for machine learning. San Mateo, CA: Morgan Kaufmann. (1993)
6. Cover, T. M., & Hart, P. E. Nearest neighbor pattern classification. IEEE Transactions on Information Theory, 13. (1967) 21-27
7. F. Aminzadeh, Future Geoscience Technology Trends in, Stratigraphic Analysis, Utilizing Advanced Geophysical, Wireline, and Borehole Technology For Petroleum Exploration and Production, GCSEPFM pp 1-6, (1996)
8. Zhongzhi Shi. MSMiner: Data Mining Platform, Keynote speech, ICMLC2002, 2002
9. Zhongzhi Shi. OKPS: Expert System Developing Tool, Technical Report, ICT of CAS, 2004
10. William W. Cohen. Fast Effective Rule Induction. In Machine Learning: Proceedings of the Twelfth International Conference, Lake Taho, California, 1995.
11. Clifford A. Brunk and Michael J. Pazzani. An investigation of noise-tolerant relational concept learning algorithms. In Proceedings of the 8th InternationalWorkshop on Machine Learning, pages 389–393, Evanston, Illinois, 1991.