Rule Induction Algorithms for Application to Geological and Petrophysical Data: Methodology

Y. Xie and C. V. Deutsch University of Alberta
(ylxie@civil.ualberta.ca, cdeutsch@civil.ualberta.ca)

A. Stan Cullick, Mobil Oil Corporation
(A_stan_cullick@mobil.com)

Abstract

This report introduces an algorithm for rule induction intended to provide new insights, improve the reliability and expedite the utilization of large geologic and petrophysical databases. Geoscientists are often challenged to predict subsurface lithologies and properties from multivariate relationships within large databases of core, wireline, and seismic data. Many data analysis techniques are used including histograms, regression, N*-dimensional histograms, cluster analysis, discrimininant anlysis, and principal components analysis to assist the geoscientist. This report introduces an algorithm that seeks to discover rule-like relationships within the data that can be used to make predictions. The method is loosely derived from a data mining technology of classification and borrows from rough sets theory, which yields rules of relations between data categories that explain decision outcomes. For a database consisting of multiple data attributes that yield one or more decision outcomes, the algorithm determines the data configurations or rules and their associated coverages, accuracies, and significances. Concepts of data attribute distinguishability and importance are introduced to assess the value of the data and the outcomes to predictability. This report presents the algorithm and is followed by a companion paper that presents an example for object-based modeling.*

KEYWORDS: geostatistics, data mining, seismic data, multivariate statistics

Introduction

Very large geological, geophysical, and petrophysical databases often contain multiple data types which must be interpreted for applications in subsurface modeling. Significant advances in discovering complex and even nonintuitive data relationships could lead to better predictions. Obvious applications include (1) predicting reservoir facies from multiple wireline logs, (2) predicting reservoir facies from multiple seismic attributes, (3) predicting stratigraphic geometries and spatial architecture from quanitified analog, outcrop, seismic, and numerical stratigraphy data, and (4) predicting reservoir permeability from wireline logs.

There are a litany of data analysis techniques used today. These include cluster analysis, principal components analysis, discriminant analysis, parametric and nonparametric regression, and *N*- dimensional histograms. Regression techniques, which include neural networks, have in common their multivariate combination of predictor variables. These combination relations are often excellent at interpolating within the data boundaries of the training data, but may be poor for extrapolation because the regression fits the training data but does not lead to an understanding of the underlying relationships in the variables. In seeking an alternative method, we identified the technology of *Rough Sets*, which holds the promise of finding data category relationships and expressing those in a rule-based language. We initiated research on rule induction to address the problem of estimating

reservoir flow unit geology from known analog data. As a first step, we have written an algorithm for rule induction.

This report describes the proposed rule induction methodology in detail. A companion paper presents an example of induction of rules between effective permeability and geometric features of object facies models. A short description of facies assignment from rules that identify reservoir facies from a wireline log suite in uncored wells is described in Appendix A. A detailed description of the implementation of the proposed rule induction algorithm can be found in Appendix B.

Background

The field of *data mining* [2, 7, 8, 10, 16, 17] has grown in recent years to deal with large databases available in different industries, in particular, in the financial and medical fields. Data mining is the identification or discovery of patterns in data. There are several different types of data mining. These include classification, clustering, association, and sequence discovery. The main focus of classification is supervised induction, that is, inference of rules and relationships from large databases. The aim is to extract knowledge from data, so that results not directly in the training data set can be predicted. The training data helps to distinguish predefined classes. Neural networks [3, 9, 12, 13, 26], decision trees [5, 6, 15, 18, 25, 28] and if-then-else rules are classification techniques. A disadvantage of NNs is that it is difficult to provide a good rationale for the predictions made, that is the rules are not always clear.

Data mining is an interdisciplinary field bringing together techniques from statistics, machine learning, artificial intelligence, pattern recognition, database, and visualization technologies. The methods used in data mining are not fundamentally different from older quantitative model-building techniques, but are natural extensions and generalizations of such methods. There are many applications of various data mining techniques to petroleum characterization [1, 4, 11, 14, 27]

A rule-based algorithm is intended to provide understandable rule-like relationships in the data. A rule is a prevailing quality or state. Induction is an instance of reasoning from a part to a whole. Rough Sets (Lin and Cercone [16], Orlowska [19], Pal and Skowron [20], Pawlak [21], Polkowski [22, 23, 24], Slowinski [29], Ziarko [30]) is a specialized method for inducing rules. Rules indicate the degree of association between variables, map data into predefined classes, and identify a finite set of categories or clusters to describe the data. The rules support specific tasks and are generated by repeated application of a certain technique, or more generally an algorithm, on the data. Rule based techniques present knowledge extracted from data in natural language with understandable semantics and syntax. Rough Sets is a promising method in the field of data mining. The methodology is concerned with the classification of uncertain knowledge. The essential idea of rough sets is to express uncertain knowledge through an approximation space, which is constructed as certain sets.

Many other methods, including regression analysis, assume that there is a functional form between the predictor and response variables. These smooth out variations and are difficult to apply to multivariate nonlinear responses. Discriminant analysis separates samples into groups based on relationships in the training data. The relations must be linear combinations of variables that are made explicit. *N*- dimensional histograms are used to delineate a relationship between a response and multiple predictors which preserves the uncertainty in the relation by reading the value of the response directly from the set of predictors. Principal components analysis is a popular technique to discover the source of variation within the data but again expresses that as a linear combination of multiple variables for which the reason for the combination is often not apparent. All of these techniques can "fit" the data but might not be good predictors.

Clustering is a method to partition the database into segments where each segment member shares similar qualities. Clustering techniques may include optimization algorithms to determine the maximum similarity among members within each group and a minimum similarity among members

across the groups.

Rule Induction Algorithm

In rough set theory, a data table is partitioned by condition attributes into three different spaces: a *positive* region, a *boundary* region, and a *negative* region. Rules are then induced from the positive and boundary region. Concepts of *coverage* and *accuracy* are introduced to describe the uncertainty in probabilistic rules inferred from the boundary region. In practice, every object in the data table may lead to a rule. The challenge is to identify significant rules, that is, accurate rules with high frequency of occurrence. Data mining has no a priori model, that is, does not assume a functional relationship between the data.

A feature of geological data is that most rules fall in the boundary region. Another aspect of geological data is that there are relatively few attributes from a classical data mining perspective. These two considerations were used to develop the concepts and terms used in this report. Following is a list, in alphabetic order, of key terms and concepts:

Accuracy : a measure of the closeness of a probability to 1 or 0.

Condition attribute value : a measure of the value contributed by every possible combination of conditioning attributes to a decision outcome.

Conditional probability : for each decision outcome and condition configuration pair, the fraction of occurences in the data table out of configuration coverage.

Conditional probability table : a complete table of all decision outcomes for all condition configurations. This table will be used in modeling of predicted outcomes.

Configuration coverage : the number of observations associated with each condition configuration.

Configuration space : the set of possible configurations of the classes for all the condition attributes.

Data table : table of observed decisions for each instance of the combination of conditioning attributes that includes all "clean" data. The set of conditioning attributes are analyzed through basic statistics and clustering so that class assignments can be made.

Decision coverage : number of observations associated with each decision outcome.

Decision space : the set of all possible decision outcomes in the data table.

Distinguishability : a measure of distinctiveness between decision outcomes, based on their probabilities.

Relative configuration coverage : a measure of closeness to a sufficiency, defined as a proportionality related to the configuration coverage.

Relative (decision) coverage : a measure of the proportion of each decision outcome out of the global number for that outcome.

Rule table : a table of the positive and negative rules with associated probabilities, which are filtered by significance. The rules are used to predict outcomes from condition configuration.

Significance : a measure of how an outcome is both sufficient and accurate; in practice, a combination of the relative coverage, accuracy, and decision coverage that leads to a rule.

	Cond. Attr. 1	Cond. Attr. 2	...	Cond. Attr. N-1	Cond. Attr. N	Dec. Attr.
Observation 1	0	2	...	1	2	1
Observation 2	2	1	...	1	2	1
...	...	...	...	...	...	...
Observation M-1	2	2	...	0	1	0
Observation M	1	0	...	2	2	2

Table 1: A typical data table for rule induction

Data Table

Table 1 is a schematic data table for rule induction. The table has M observations, N condition attributes and a single decision attribute. Initially, the condition attributes and decision attribute can have different values or categories, which are finally transformed into discreted codes representing the corresponding value level. The codes are based on statistical analysis of the data. For example, the data table might consist of M observations of seismic facies decision attributes, for which the continuous conditioning attributes are amplitude, phase, frequency content, velocity, and density, which have been coded in two to four classifications each.

In Table 1, M is the number of observations, N is the number of condition attributes, $n_i, (i = 1, \ldots, N)$ is the number of categorical values that condition attribute i can take, n_o is the number of categorical value the decision attribute can have. It should be pointed out that the number of categorical values for an attribute is fixed for a problem, say $n_1 = 3$, but the codes themselves can be different; both $0, 1, 2$ and $5, 9, 11$ are valid codes for the three categorical values of attribute 1.

Without loss of generality, it is assumed that there is a single decision attribute. Multiple decision attributes merely increase the number of rules and the configuration space.

Configuration and Data Coverages

The total number of possible configurations of the conditioning attributes is: $N_N = \prod_{i=1}^{N} n_i$, where $n_i, i = 1, \ldots, N$ is the number of categorical values of condition attribute i.

Note that many configurations may not appear in the training data table.

Further, define $N_i (= \prod_{k=1}^{i} n_k)$ as the cumulative number for condition attribute $i, i = 1, \ldots, N$ and all "lower" condition attributes.

Configuration index j, $j = 1, \ldots, N_N$, has correspondence with the values $l_i, i = 1, \ldots, N$ taken by each condition attribute through $j = \sum_{i=1}^{N} (l_i - 1) \times \frac{N_N}{N_i} + 1$, where l_i is the value, $l_i \in 1, \ldots, n_i$, taken by condition attribute i.

The retrival of the condition atttribute configuration l_i, $i = 1, \ldots, N$ from a configuration index j will be accomplished:

$$A_0 = N_N \text{ and } B_0 = j$$

$$A_i = \frac{A_{i-1}}{n_i}, i = 1, \ldots, N-1$$

$$l_i = int(\frac{B_{i-1} - 1}{A_i}) + 1, i = 1, \ldots, N-1$$

$$B_i = B_{i-1} - (l_i - 1) \times A_i, i = 1, \ldots, N-1$$

$$l_N = B_{N-1}$$

Every configuration is a potential rule in the system. Since the decision attribute can have n_o outcomes, there are $N_N \times n_o$ potential rules for the system.

From the data table, the occurence of configuration j with an outcome o is counted as $C_{j,o}$, ($j = 1, \ldots, N_N$ and $o = 1, \ldots, n_o$). If $C_{j,o} = 0$, then there are no observations in the data table corresponding to that configuration and outcome pair. It is noticed that the number of observations M, that is, the size of the data table is:

$$M = \sum_{j=1}^{N_N} \sum_{o=1}^{n_o} C_{j,o}$$

The *configuration coverage* C_j is the total number of observations associated with configuration j , i.e.,

$$C_j = \sum_{o=1}^{n_o} C_{j,o}, j = 1, \ldots, N$$

The *decision coverage* D_o is defined as the number of observations associated with outcome o of the condition attribute, i.e.,

$$D_o = \sum_{j=1}^{N_N} C_{j,o}, j = 1, \ldots, N$$

The *relative (configuration) coverage* of configuration j, $\hat{C}_j$ is a measure used to determine a closeness to sufficiency of number of observations in obtaining a reliable rule. $\hat{C}_j$ is asymptotic to one, as the coverage increases in number for configuration j.

$$\hat{C}_j = 1 - \frac{1}{C_j}$$

The definition of *relative (configuration) coverage* is not unique. For example, if a prori information exists to tell us that the number of observations is sufficient when exceeding a critical number is avialable, the *relative coverage* can be defined as:

$$C_j = 1 - exp^{\frac{-C_j}{n^{crit}}}$$

where n^{crit} is a critical number for the sufficiency of the number of observations.

The *relative (decision) coverage* of a decision value, $\hat{D}_{j,o}$, is a measure of the proportion of each decision outcome out of the global number for that outcome, i.e.,

$$\hat{D}_{j,o} = \frac{C_{j,o}}{D_o}$$

This will be used to place more weight on decision values with small global proportions to avoid cases where $C_{j,o}$ is small, but is close to D_o. In such a case, this configuration is very important to that decision.

Accuracy

The conditional probability of each outcome value, $o, (o = 1, \ldots, n_o)$ of configuration j is defined as:

$$P_{o|j} = \frac{C_{j,o}}{C_j}$$

Since there are n_o possible decision attributes, a probability of $\frac{1}{n_o}$ implies no information. Any conditional probability different from $\frac{1}{n_o}$ entails preference in the decision category, i.e., hints at knowledge. Specifically, as $P_{o|j}$ nears 1, configuration j implies decision category o. A closeness of $P_{o|j}$ to 0 means configuration j does not lead to decision category o. Conditional probabilities close to 1 or 0 contain equally important information for rule induction. The former leads to a positive rule relating configuration j to a specific decision category o and the latter leads to a negative rule relating configuration to some other decision category. This leads us to define a measure of accuracy as follows:

$$a_{o|j} = \begin{cases} p_{o|j} \times n_o - 1 & \text{if } p_{o|j} \leq \frac{1}{n_o} \\ \frac{(p_{o|j} \times n_o)}{n_o - 1} & \text{otherwise} \end{cases}$$

The accuracy $a_{o|j} = 0$ if $P_{o|j} = \frac{1}{n_o}$ (no knowledge); the accuracy $a_{o|j} = 1$ when we have $P_{o|j}$ equals to 1 (perfect positive rule); and, $a_{o|j} = 11$ when we have $P_{o|j}$ equals to 0 (perfect negative rule). A graph of accuracy versus probability is shown on Figure 1.

Significance

The *significance* is defined as a combined measure of the *accuracy* and *coverage*, which is used to provide a relative ranking to the rules. The significance is the product of the *accuracy* and the *relative configuration coverage*, i.e.,

$$S_{j,o} = \hat{C}_j \times a_{o|j}$$

This equation holds for negative accuracy. For positive accuracy, the significance is further modified to account for the *relative decision coverage*, i.e.,

$$S_{j,o} = \hat{C}_j \times a_{o|j} \times \hat{D}_{j,o}$$

Rule Table

Positive rules are extracted from the configurations ranked from positive 1 to zero by positive significances, and negative rules are extracted from the configurations ranked from negative one to zero negative significances. A practioner can decide whether to assign a minimum value for positive significance for positive rules and a maximum (negative) value for negative significance for negative rules.

The rules are those configurations that meet the assigned significance criteria, whether positive or negative. Positive rules provide evidence for the decision outcome for a rule (configuration) and negative rules provide evidence against the decision that meet the cutoff criteria, respectively. The degree of positive evidence is taken as the positive value of the accuracy and the negative evidence is taken as the negative accuracy. If a decision outcome's accuracy does not meet the significance criteria, the rule table entry is simple *NA* for not sufficient evidence for supporting that outcome, either positive or negative.

It is expected that the rule tables, with $a_{o|j}$, $p_{o|j}$ or $s_{o|j}$, will have many *NA* entries for rule induction with a real data set due to both the lack of any coverage and insignificance of some configurations. We will describe the procedure to eliminate or minimize the number of *NA* entries based on the maximal retrival of information from the data set by searching compatible rules in the rule set generated from subsets.

Distinguishability

In a data mining exercise, there is often an implicit assumption that all the decision outcome attributes are important in the context of the data, that is, they are distinguishable. This assumption

Observation (m=)	Predicted probability from the rules	Real outcome ($\tilde{o}_m$)
1	$p_{1,1}, p_{1,2}, \ldots, p_{1,n_o}$	$\tilde{o}_1$
2	$p_{2,1}, p_{2,2}, \ldots, p_{2,n_o}$	$\tilde{o}_2$
...	...	...
m	$p_{m,1}, p_{m,2}, \ldots, p_{m,n_o}$	$\tilde{o}_m$
...	...	...
$M-1$	$p_{M-1,1}, p_{M-1,2}, \ldots, p_{M-1,n_o}$	$\tilde{o}_{M-1}$
M	$p_{M,1}, p_{M,2}, \ldots, p_{M,n_o}$	$\tilde{o}_M$

Table 2: Predicted Probability from the Rules

is not always true and a measure of "distinctiveness" can assist in determining which decision outcomes may not be distinct on the basis of the data table.

The conditional probability of the occurrence of decision category o given configuration j of condition attributes is the essence of measures such as *accuracy* and *significance*. In rule induction and classification, such conditional probabilities play a critical role. For optimal classification the conditional probabilities $p_{o|j}$ should be maximum for every observation where the decision category is actually o. Therefore, the average of $p_{o|j}$ should be a proper measure of the quality of the rules and its classification.

Once the rule induction computation (training) is finished, one can construct a data table like the one shown in Table 2. For each observation in the training data set, predicted probability will be assigned to it by looking for the configuration j the observation corresponds to and finding the corresponding $p_{o|j}$.

The data table may be divided into n_o classes with an indicator function:

$$ind(m;o) = \begin{cases} 1 & \text{if } \tilde{o}_m = o \\ 0 & \text{otherwise} \end{cases}$$

The summation of $ind(m;o)$ is a counting of all outcomes related to decision value o, that actually is the *decision coverage*:

$$D_o = \sum_{m=1}^{M} ind(m;o)$$

The average probability associated with prediction of o is:

$$\hat{p_o} = \frac{1}{D_o} \sum_{m=1}^{M} ind(m;o) \times p_{m,\tilde{o}_m}, o = 1, \ldots, n_o$$

As $\hat{p_o}$ approaches 1, prediction should be good. As $\hat{p_o}$ goes to $p_o = \frac{D_o}{M}$, i.e., the global proportion for decision outcome o, there is no information provided by the rules.

Therefore, the relative information value for prediction outcome o is defined as:

$$I_o = \frac{\hat{p}_o - p_o}{p_o}$$

A single information value could be defined as the average over all outcomes $o = 1, \ldots, n_o$, i.e.,

$$\hat{I} = \frac{1}{n_o} \sum_{o=1}^{n_o} I_o$$

The expected probability and information measure can be used to determine whether outcomes o and o' are distinct.

Number of condition attribute	Number of subsets with such number of condition attributes
N	*1*
N-1	*N*
...	...
l	$C_N^l = \frac{N!}{(N-l)!l!}$
...	...
1	*N*

Table 3: Number of subsets with various number of condition attributes

Let's consider merging two different decision outcomes *o* and *o'*, then we have $n_o - 1$ new decision outcomes and the predicted probability table as Table 2 will be updated. Updating the predicted probability is straighforward. We derive $n_o - 1$ probability values from the previous n_o values. One of them will be $p_{o|j} + p_{o'|j}$ when the actual outcome turns out to be *o* or *o'*.

Following the same procedure calculating $\hat{p}$ and I above, new $\frac{D_o}{P_o}$, the P_o, and I_o, and the $\hat{I}$ values will be calculated. If we define $\hat{I}_{o,o'}$ as the information value with *o* and *o'* merged, the change in $\hat{I}$ is:

$$\triangle \hat{I}_{o:o'} = \hat{I}_{o:o'} - \hat{I}$$

$\triangle \hat{I}_{o:o'} > 0$ means that *o* and *o'* merged leads to an improved prediction. Thus, one should consider whether to treat these outcomes together, i.e., as indistinguishable, since the given training data table cannot differentiate them. One could seek to obtain additional data on a different data type that would distinguish those outcomes. $\triangle \hat{I}_{o:o'} < 0$ means that merging the outcomes leads to a poorer prediction. Thus, the outcomes should be treated as distinct.

Lumping could be considered for all possible pairs of outcome values, and the $\triangle \hat{I}_s$ can be tabulated, plotted or ranked from high to low. The high values are candidates for merging while the low values should be kept separate. The $\triangle \hat{I}_{o:o'}$ table can be visualized as a color map. For example, Figure 2 shows a map of the matrix pairing of nine decision outcomes' $\triangle \hat{I}$ values. The diagonal is the pairing of each outcome with itself and has a $\triangle \hat{I}$ value of 0. The warm colors indicate a positive $\triangle \hat{I}$ value which shows high distinguishability of outcomes. The rows of red and orange for outcomes 7, 8, 9 indicate that those are very distinguishable from other outcomes. The upper left corner area with cool colors of negative $\triangle \hat{I}$ values indicate that outcomes 1-2, 2-3,and 2-4 are not as distinguishable and might be candidates for some combination.

Data or Condition Attribute Value

The same measure of information can be used to consider the value (importance) of condition attributes or sets of condition attributes. All procedures described above are based on the entire data set with all condition attributes, but the procedures can be applied to any data subset with partial condition attributes.

For a *N* condition attribute data set, there are $2^N - 1 (= \sum_{l=1}^{N} C_N^l)$, ($l$ is the number of condition attribute in the subset) subsets. The number of subsets is shown in Table 3.

For each subset of condition attibutes, a new data set is derived and the algorithm is applied. A new set of configurations of condition attributes are obtained and $p_{o|j}$s are calculated for every decision category *o* for a given configuration *j*. For each individual decision category, the average conditional probability of the occurrence of decision category *o* when given the data entry having decision category *o* is also calculated as well as the information measure.

The information value for each subset is, for example, $\hat{I}_{\{1,2,3,...,N\}} = \hat{I}$ for all attributes, $\hat{I}_{\{2,3,...,N\}}$ for leaving out condition attribute 1, $\hat{I}_{\{2\}}$ for just attribute 2, *etc.* These results can be plotted as

Number of condition attribute	1	2	...	*N-1*
highest value information value	2	2_3	...	2_4_5_(*N-2*)
...	...	...	...	...
lowest values information value	*N-5*	1_4	...	1_2_4_(*N*)

Table 4: Number of subsets with various number of condition attributes

a graph or a table that ranks the value of the attributes. For example, Table 4 shows a ranking by condition attribute subsets schematically.

The largest values, or maximums of information in the subsets with the same number of condition attributes, would normally trend to increase as the number of attributes in the subsets increases. However, noisy data or attributes that do not yield additional information value could reduce $\hat{I}$.

Probability table

The utilization of the rules is two-fold: (1) gain insight and make general observations from the rule table, and (2) use the conditional probability associated with each configuration and decision outcome in the probability table to model the predictions probabilistically. These application will be illustrated in the companion paper.

The rule (probability) tables built from the entire data set are not complete since there may be no observations for some condition attribute configurations and there are some rules have been screened out due to small significance. A complete rule table is necessary for practical usage of rule induction algorithm and empty entries in the rule table must be filled in.

Filling blank entries in the rule table will be accomplished based on significant rule derived from data subsets. In the rule set induced from a data subset, we search for the rules with the configurations which are the subset of the configurations where we have not got rules for them from the entire data set. If rules are significant they will be used as in the rule table.

There are many subsets of condition attributes, therefore, it is necessary to order the subsets for the searching and filling procedure. This ordering will be based on their information value. The ordered subsets will be considered until all blanks are filled in. Those entries filled from subsets searching will be flagged to indicate rules from which subsets are filled.

The predictions can then utilize the probabilities in modeling. There is opportunity to reduce the uncertainties by incorporating additional data sources in the modeling phase.

Implementation

An implementation issue is the selection of categorical classes by application of thresholds to a continuous variable. This could be performed by optimization.

Regarding the rule induction implementation there are issues related to (1) the physical meaning of the rules, (2) the usage of results, (3) the meaning of a probability prediction, (4) data integration, and (5) subjective understanding. The significance is between -1 to + 1. In principle, the potential rules with largest absolute significance values will be chosen as important rules, but there exists ambiguity of how to define big and small, e.g., is 0.9 close enough? When there are -1 and/or +1 in the significance list, those rules will be chosen without question. The others will depend on judgment.

Fortunately, we do not care much about rules in the middle of the list. The rule induction is only as good as the training data. But in contrast to other techniques, e.g. neural networks, each

rule should be clear and understandable to a practioner as to its applicability. Limited to training data, representativity, "meaning" of probability value or significance value (nonlinear).

Summary of Procedure

The rule induction procedure consists of the following steps:

- Prepare the data table and determine optimal classes for the conditioning attributes (discretization).
- Enumerate all possible configurations and outcomes.
- Calculate the *configuration coverage, decision coverage, accuracy*, and *significance*.
- Present a table for each outcome $o = 1, \ldots, n_o$ with the configurations sorted from low to high *significance*.
- Apply cutoff criteria on significance to build a Rule table.
- Evaluate information value for the condition attributes and/or decision outcome
- Build up a probability table that picks up probabilities from subsets of the conditioning attributes that meet significance criteria.

Discussion

The proposed algorithm and significance measures work well for the simple example. The most important positive and negative rules are retrieved successfully. The proposed significance definition combines measures of accuracy and coverage and serves as a "quality" measure of rules. Also, the significance identifies positive and negative rules, similar to the positive region and negative region in rough sets.

The proposed rule induction technique is suited to geological data where most attributes are significant. The proposed significance measure can be used in combination with other rule induction techniques and serves as a ranking measure to identify the most important rules. In general, however, the algorithm will need to be extended to include attribute reduction.

Careful examination of a data table may lead an experienced person to infer similar, if not the same rules; however, there are many advantages to automatic rule induction. The procedure works for very large data tables with many attributes, it is repeatable, and avoids personal biases. It could also lead to nonintuitive, but meaningful data relations. The effort in making predictions can be greatly reduced.

Appendix A and the companion paper will present petrophysical examples and introduce next steps in the research.

Appendix A: Facies Assignment from Wireline Well Logs

This example consists of various well logs data and cored facies assignment. The purpose of the study is to investigate the potential to use well logs for facies assignment in uncored wells.

Various well logs data are available including *Gamma Ray (GR)*, *Resistivity (Res)*, *Bulk Density (BD)* and *Neutron Activation (NA)*. Both *Bulk Density (BD)* and *Neutron Activation (NA)* are measures of density and the cross over of the difference between `normalized` *Bulk Density (BD)* and `normalized` *Neutron Activation* is important in well log interpretation. Therefore, instead of using *Neutron Activation*, the difference of `normalized` *Bulk Density (BD)* and *Neutron Activation*

Facies	Description	Number of data
1	Cross-bedded med. to coarse-grained sandstone	158
2	Medium to fine-grained sandstone	283
3	Well sorted fine-grained sandstone; Shoreface	241
4	Very fine to fine-grained sandstone	181
5	Heterolithic: coarse-/fine-grained intervals; sand-filled burrow	117
6	Transition zone, lower shoreface to inner shelf	77
7	Muddy v. fine-grained sandstone and sandy shale	62
8	Very fine sandy shale	70
9	Shale	27

Table 5: Description of facies

Code	0	1	2	3	4	5	6	7	8
GR	0-30	30-55	≥ 55						
Resistivity	0-1.5	1.5-2.5	2.5-20	≥ 20					
Bulk Density	0-2.05	2.05-2.30	≥ 2.30						
Crossover	+	-0.5-0	≤ -0.5						
Water/Oil Zone	0 (Water Zone)	1 (Oil Zone)	2 (Gas Zone)						
Facies	1	2	3	4	5	6	7	8	9

Table 6: Coding of condition and decision attributes

are calculated as a new attribute, *CROSSOVER*, which is used instead of *Neutron Activation*. Also, we have water/oil/gas zone information for the wells and we use this piece of information as well.

There are 9 facies recognized from cores and Table 5 lists the description of facies. The depth resolution in the wells is 0.5 feet. After removing missing data, there are a total of 1216 data points (observations) and the number of observations for each facies (decision value) are listed in Table 5 as well. Figure 3 shows the distribution of the four condition attributes of the entire data set.

In order to use the rule induction algorithm, the four continuous attributes need to be discretized. This is not a trival issue and research is ongoing for optimal disrectization. We inspected and compared the distributions of the four attributes for each individual facies (decision attribute) to decide the discretization. The discretizations of the attributes are listed in Table 6. Note that we number all the attributes from 0. Finally a 1216 by 5 data table is constructed for rule induction.

Theoretically there are 324 ($= 3\times4\times3\times3\times3$) configurations of condition attributes; however, only 54 configurations have nonzero coverage. There are 31 ($= 2^5 - 1$) subsets of condition attributes and the complete rule table are obtained from the entire attribute subset supplemented with blank entries filled with rules from the subsets. The top plot of Figure 4 shows the ranked information measure of all 31 subsets of condition attributes. The bottom plot of Figure 4 shows the ranked information measure within groups of subsets with the same number of condition attributes. Even though there is no universal increase in the information measure as the increase of number of condition attributes, the largest information in the subsets with the same number of condition attributes increases as the number of condition attributes in the subsets increases. For this example, the entire set has the largest information measure, therefore the entire set of variables should be used for rule induction.

The complete set of rules and probability table contain 324 lines (configurations). In each line, the positive numbers are probability values for decision outcome with a significant positive *significance* and the negative number are the probabilities for decision outcome with a significant negative *significance* added a minus sign. Table 7 lists several lines of the final rule table filtered by the significance threshold and Table 8 lists the same lines in the probability table. The difference

No	Code of Condition Attributes					Filtered Significance of Decision Values									a	b	c
1	0	0	0	0	0	.00	.00	.93	.00	.00	.00	.00	.00	.00	3	23	15
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
66	0	2	1	0	2	-1.00	-1.00	-1.00	.00	.00	.00	.00	.22	.22	4	30	36
67	0	2	1	1	0	1.00	.00	.00	.00	.00	.00	.00	.00	.00	5	31	8
68	0	2	1	1	1	-1.00	.00	.22	.39	-1.00	.22	-.99	.00	.00	4	30	99
69	0	2	1	1	2	.00	.00	-1.00	.00	.58	.17	.00	.00	-.99	4	30	98
70	0	2	1	2	0	1.00	.00	.00	.00	.00	.00	.00	.00	.00	4	27	8
71	0	2	1	2	1	-1.00	.00	.00	.36	-1.00	.22	.00	.00	.00	3	23	122
72	0	2	1	2	2	.00	.00	.00	.00	1.00	.00	.00	.00	.00	4	30	5
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
324	2	3	2	2	2	.00	.00	.00	.00	.00	.00	.00	.00	.00	4	28	11

Table 7: Rules in the complete rule table a: Level of subset b: Index of subset c: Configuration coverage

No	Code of Condition Attributes					Conditional Probability of Decision Values									a	b	c
1	0	0	0	0	0	.00	.00	.93	.07	.00	.00	.00	.00	.00	3	23	15
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
66	0	2	1	0	2	.00	.00	.00	.02	.14	.19	.19	.22	.22	4	30	36
67	0	2	1	1	0	1.00	.00	.00	.00	.00	.00	.00	.00	.00	5	31	8
68	0	2	1	1	1	.00	.11	.22	.39	.00	.22	.01	.02	.02	4	30	99
69	0	2	1	1	2	.03	.06	.00	.06	.58	.17	.06	.02	.01	4	30	98
70	0	2	1	2	0	1.00	.00	.00	.00	.00	.00	.00	.00	.00	4	27	8
71	0	2	1	2	1	.00	.09	.18	.36	.00	.22	.12	.02	.02	3	23	122
72	0	2	1	2	2	.00	.00	.00	.00	1.00	.00	.00	.00	.00	4	30	5
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
324	2	3	2	2	2	.09	.27	.00	.27	.36	.00	.00	.00	.00	4	28	11

Table 8: Conditional Probability in the complete probability table a: Level of subset b: Index of subset c: Configuration coverage

between these two tables are the effects of the significance thresholds used for filtering the rules.

The rules were applied to the training data set. Figure 5 presents the rules when applied to the training data in Well 1. The top plot shows the positive rules which are the probabilities of facies appearing along the depth and the bottom plot shows the negative rules which are the probabilities of facies unlikely to appear.

Figure 6 shows the true cored facies of the training data in Well one (top) and the assigned facies based on well logs (bottom). For this example, the information changes when *lumpping* decision outcome pairewise are shown in Figure 2.

Appendix B: Description of Program: `ruleind`

Figure 7 shows the parameter file of `ruleind`. Figure 8 displays a flow chart of the algorithm.

References

[1] R. S. Balch, B. S. Stubbs, W. W. Weiss, and S. Wo. Using artificial intelligence to correct multiple seismic attributes to reservoir properties. In *SPE Annual Tech Conference and Exhibition*, Houston, TX, October 5-8 1999. SPE 56733.

[2] M. J. A. Berry and G. Linoff. *Data Mining Techniques, For Marketing, Sales, and Customer Support*. Wiley Computer Publishing, John Wiley & Sons Inc., New York, 1997.

[3] C. M. Bishop. *Neural Networks for Pattern Recognition*. Oxford University Press, New York, 1995.

[4] S. Bradley, Usama Fayyad, and O. L. Mangasarian. Mathematical programming for data mining: formulations and challenges. *INORMS J. on Computing*, 11:217–238, 1999.

[5] L. Breiman, J. Friedman, R. Olshen, and C. Stone. *Classification and Regression Tree*. Chapman and Hall, New York, 1984.

[6] W. Buntine. Learning classification trees. *Statistics and Computing*, pages 63–73, 1992.

[7] P. Cabena, P. Hadjinian, R. Stadler, J. Verhees, and A. Zanasi. *Discovering Data Mining from Concept to Implementation*. Prentice Hall, New York, 1997.

[8] Two Crows Corporation. Introduction to data mining and knowledge discovery. 1999.

[9] S. P. Curram and J. Mingers. Neural networks, decision tree induction and discriminant analysis: an empirical comparison. *Journal of the Operational Research Society*, 45:440–450, 1994.

[10] U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy. *Advances in Knowledge Discovery and Data Mining*. MIT Press, Cambridge, MA, 1996.

[11] J. R. Hook, J. A. Nieto, C. T. Kalkomey, and D. Ellis. Facies and permeability prediction from wireline logs and core - a north sea case study. In *35th Annual SPWLA Logging Symposium*, Tulsa, 1994.

[12] T. Kohonen. *Self-Organization and Association Memory (3rd edition)*. Springer-Verlag, 1989.

[13] T. Kohonen. *Self-Organization Maps*. Springer-Verlag, Heidelberg, 1995.

[14] Sang Heon Lee and Akhil Datta-Gupta. Electrofacies characterization and permeability predictions in carbonate reservoirs: roles of multivariate analysis and nonparametric regression. In *SPE Annual Tech Conference and Exhibition*, Houston, TX, October 5-8 1999. SPE 56658.

[15] T. S. Lim, W. Y. Loh, and Y. S. Shih. A comparison of prediction accuracy, complexity, and training time of thirty-three old and new classification algorithms. *The Machine Learning Journal*, pages 1–27, 1999.

[16] T. Y. Lin and N. Cercone. *Rough Sets and Data Mining: Analysis for Imprecise Data*. Kluwer Academic Publisher, 1998.

[17] Syam Menon and Ramesh Sharda. Data mining update: new modes to persue old objectives. *ORMS Today*, pages 26–29, June 1999.

[18] S. K. Murthy, S. Kasif, and S. Salzberg. A system for induction of oblique decision trees. *Journal of Artificial Intelligence Research*, 2:1–33, 1994.

[19] Ewa Orlowska. Incomplete information: Rough set analysis. In *Vol 13, Studies in Fuzziness and soft computing*, page 620. Physica Verlag, 1998.

[20] S.K. Pal and A. Skowron. *Rough Fuzzy Hybridization: A New Trend in Decision-Making*. Springer Verlag, New York, 1999.

[21] Zdzislaw Pawlak. *Rough Sets: Theoretical Aspects of Reasoning About Data*. Kluwer Academic Publisher, Dordrecht, 1991.

[22] L. Polkowski and A. Skowron. Rough sets and current trends in computing. In *First international Conference, RSCTC'98*, page 601, Warsaw, Poland, 1998. Springer.

[23] L. Polkowski and A. Skowron. Rough sets in knowledge discovery i: Methodology and applications. In *Vol 18, Studies in Fuzziness and soft computing*, page 570. Physica Verlag, 1998.

[24] L. Polkowski and A. Skowron. Rough sets in knowledge discovery ii: Applications, case studies, and software systems. In *Vol 19, Studies in Fuzziness and soft computing*, page 601. Physica Verlag, 1998.

[25] J. R. Quinlan. Induction of decision tree. *Machine Learning*, 1:81–106, 1986.

[26] B. D. Ripley. *Pattern Recognition and Neural Networks*. Cambridge University Press, Cambridge, 1996.

[27] J. R. Scheevel and K. Payrazyan. Principal component analysis applied to 3d seismic data for reservoir property estimation. In *SPE Annual Tech Conference and Exhibition*, Houston, TX, October 5-8 1999. SPE 56734.

[28] Y. S. Shih. Families of splitting criteria for classification trees. *Statistics and Computing*, 1999.

[29] Roman Slowinski and Roman Slowinski. *Intelligent Decision Support: Handbook of Applications and Advances of the Rough Sets Theorya*. Kluwer Academic Publisher, Dordrecht, 1992.

[30] W. P. Ziarko. Rough sets, fuzzy sets, and knowledge discovery. In *Proceedings of the International Workshop on Rough Sets and Knowledge Discovery (RSKD '93)*, Banff, Alberta, Canada, 12-15 October 1993. Springer-Verlag.

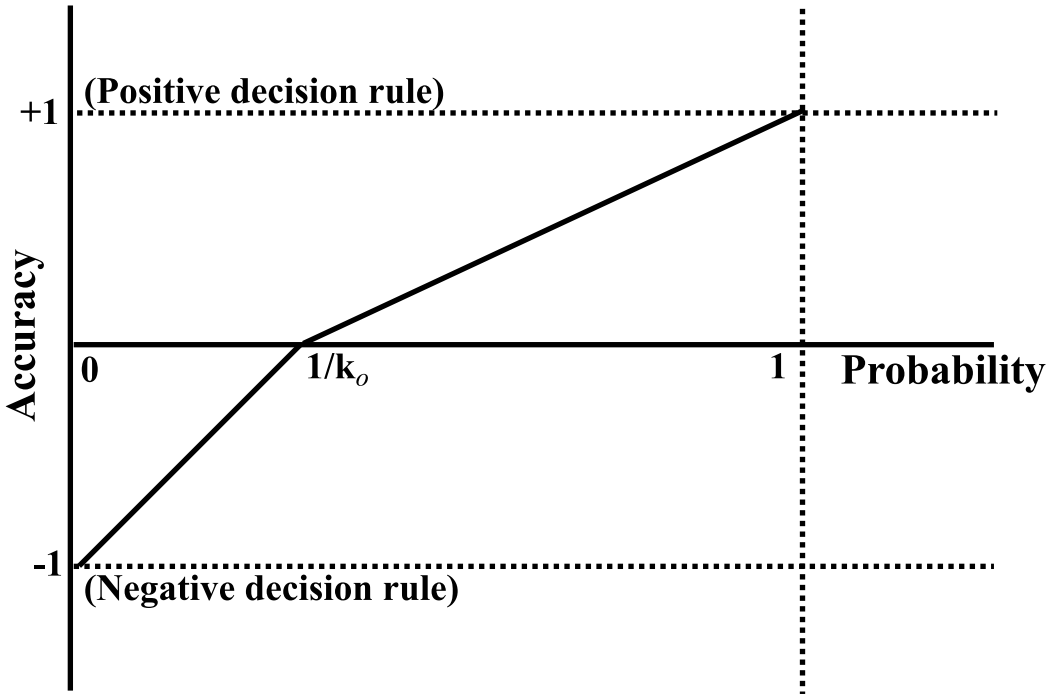

Figure 1: Relationship between accuracy and conditional probability

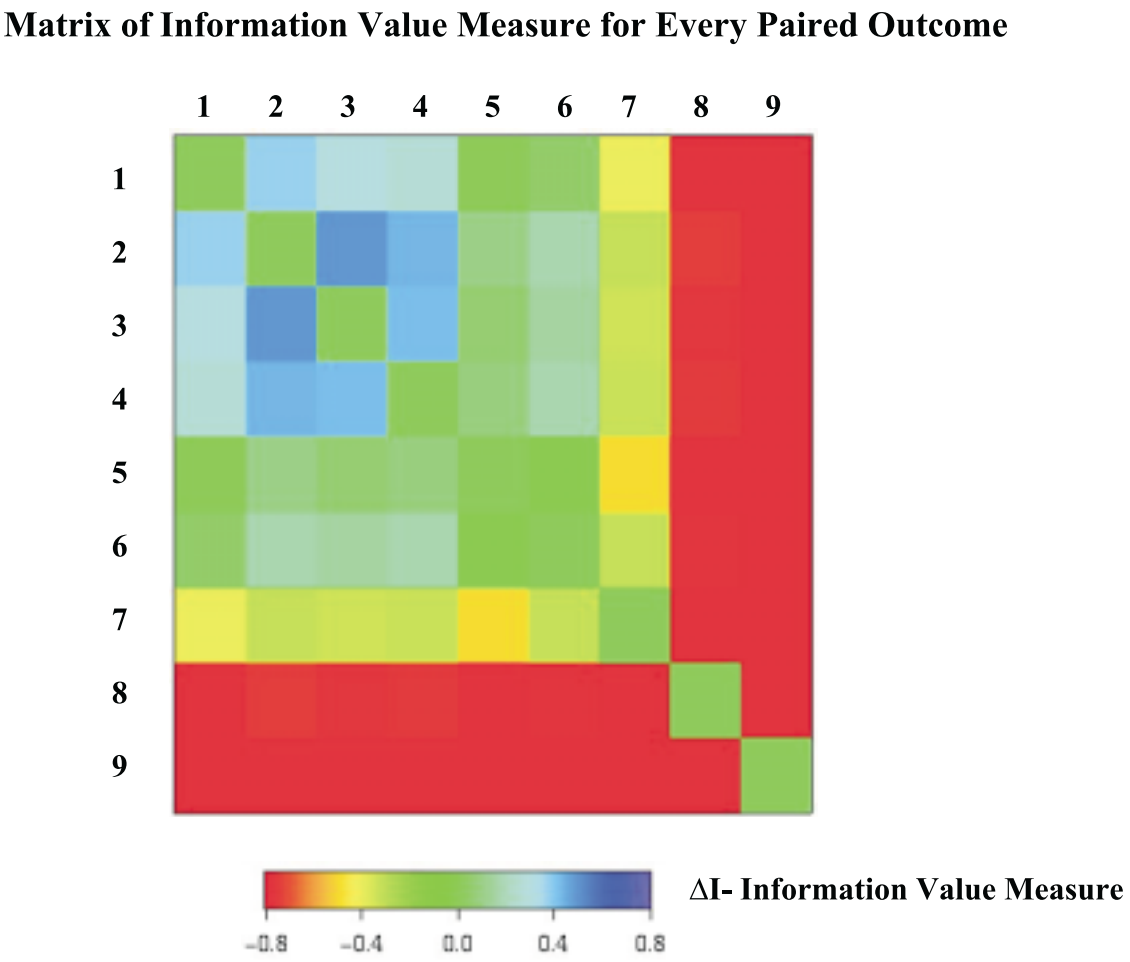

Figure 2: Changes in information measure when decision values lumped

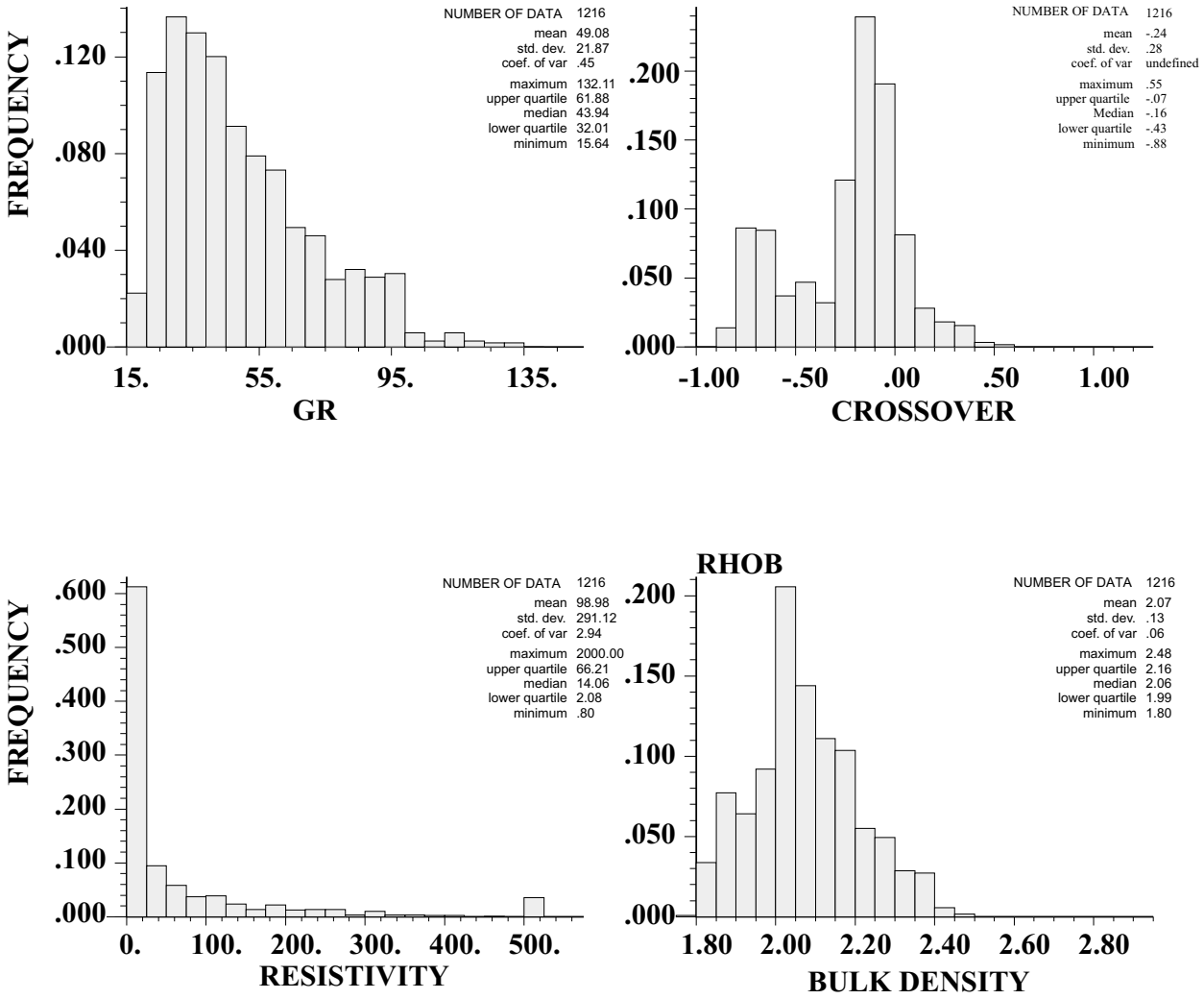

Figure 3: Hisograms of Condition Attributes

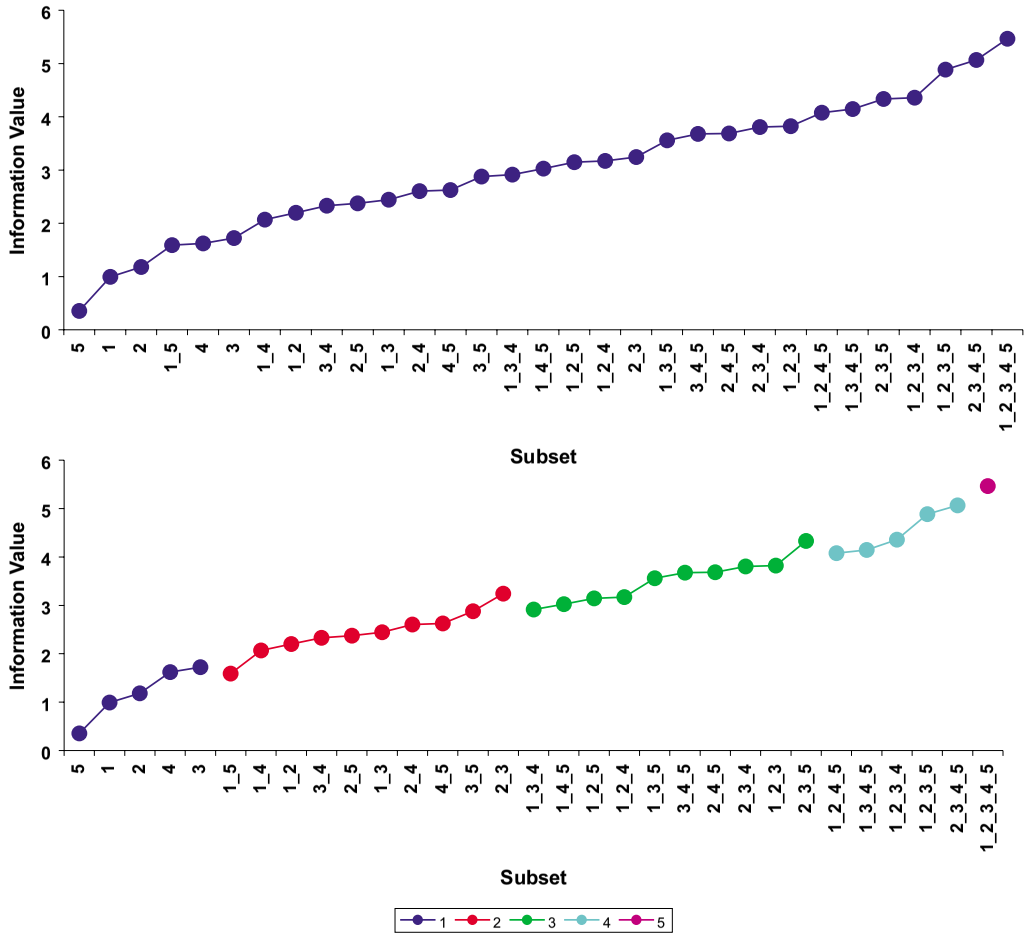

Figure 4: Information value of all subsets

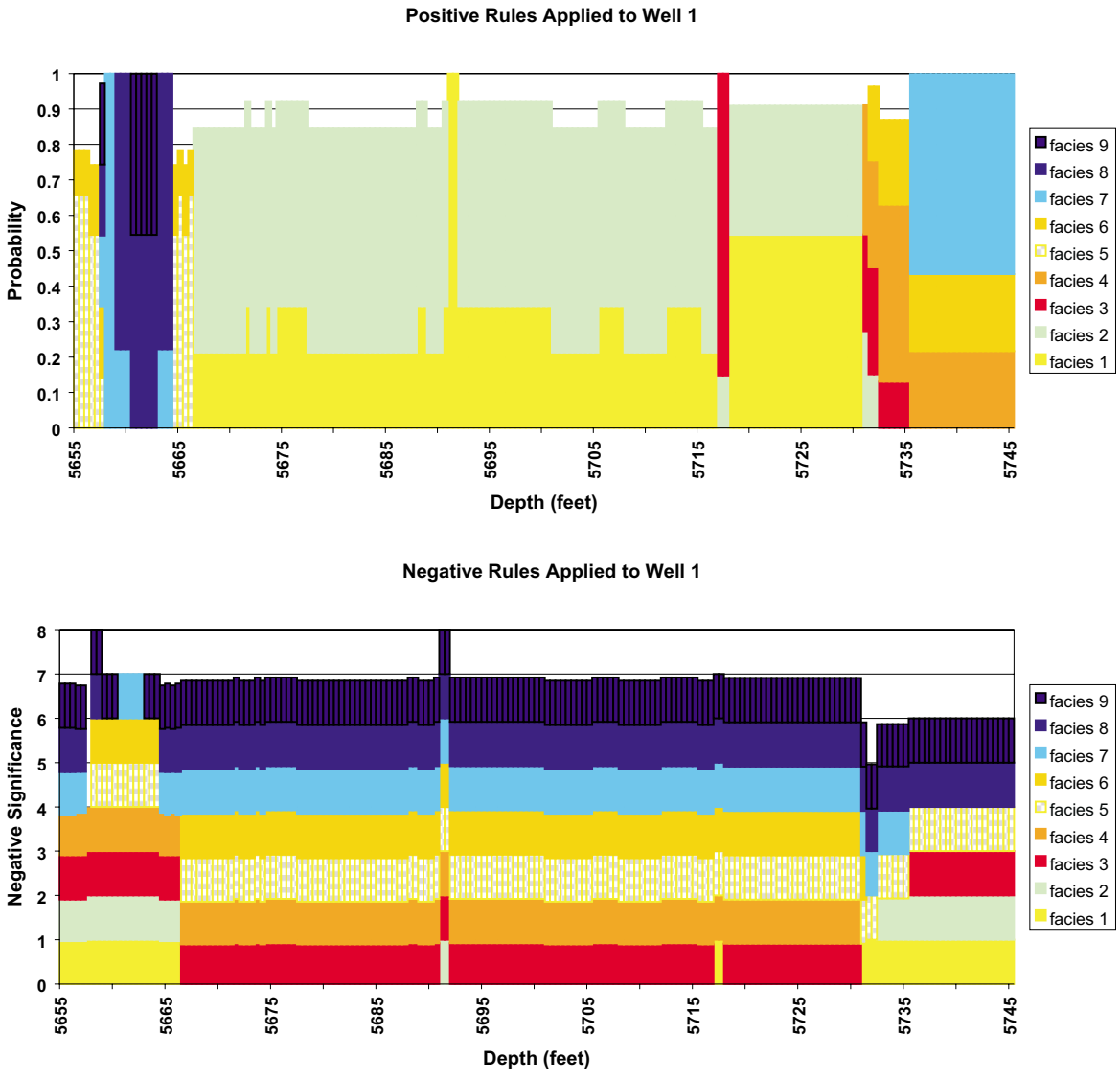

Figure 5: Induced rules (*top*: positive and *bottom* negative) when applied to training data from one well

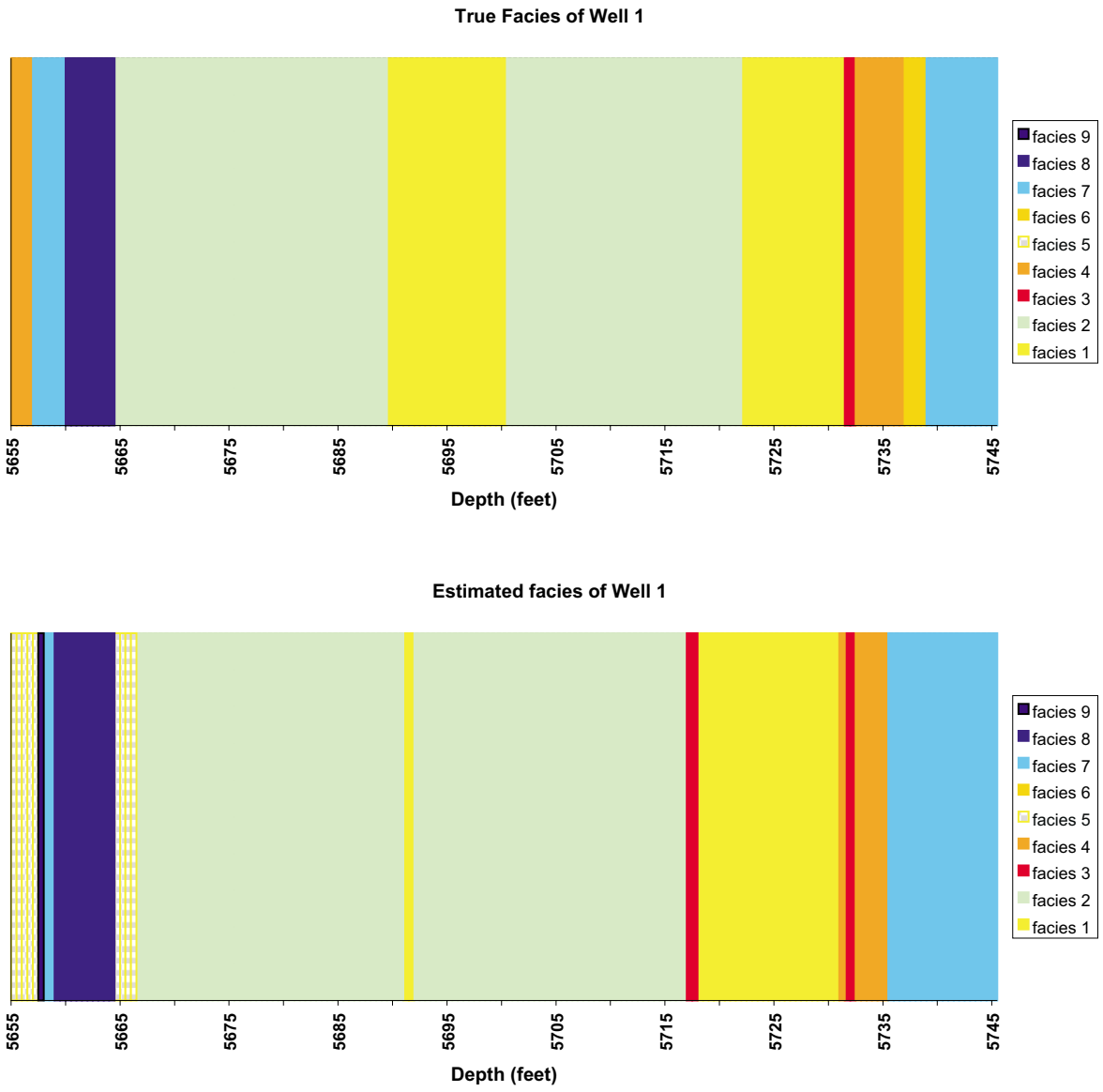

Figure 6: *top*: *true* cored facies *bottom*: assigned facies based on rules and well logs

```
                Parameters for ruleind
                **********************

START OF PARAMETERS:
facies.sys                              - input data file for rule induction
5                                       - no. of cond. attr.
1,2,3,4,5                               - cols. of cond. attr.
3,4,3,3,3                               - no. of levels of cond. attrs.
0,1,2                                   - levels of cond. attr. 1
0,1,2,3                                 - levels of cond. attr. 2
0,1,2                                   - levels of cond. attr. 3
0,1,2                                   - levels of cond. attr. 4
0,1,2                                   - levels of cond. attr. 5
1                                       - no. of decision attr.
6                                       - col of decision attr.
9                                       - no. of levels of decision attr.
0,1,2,3,4,5,6,7,8                       - levels of decision attr.
facies                                  - name of project
```

Figure 7: Parameter file of `ruleind`

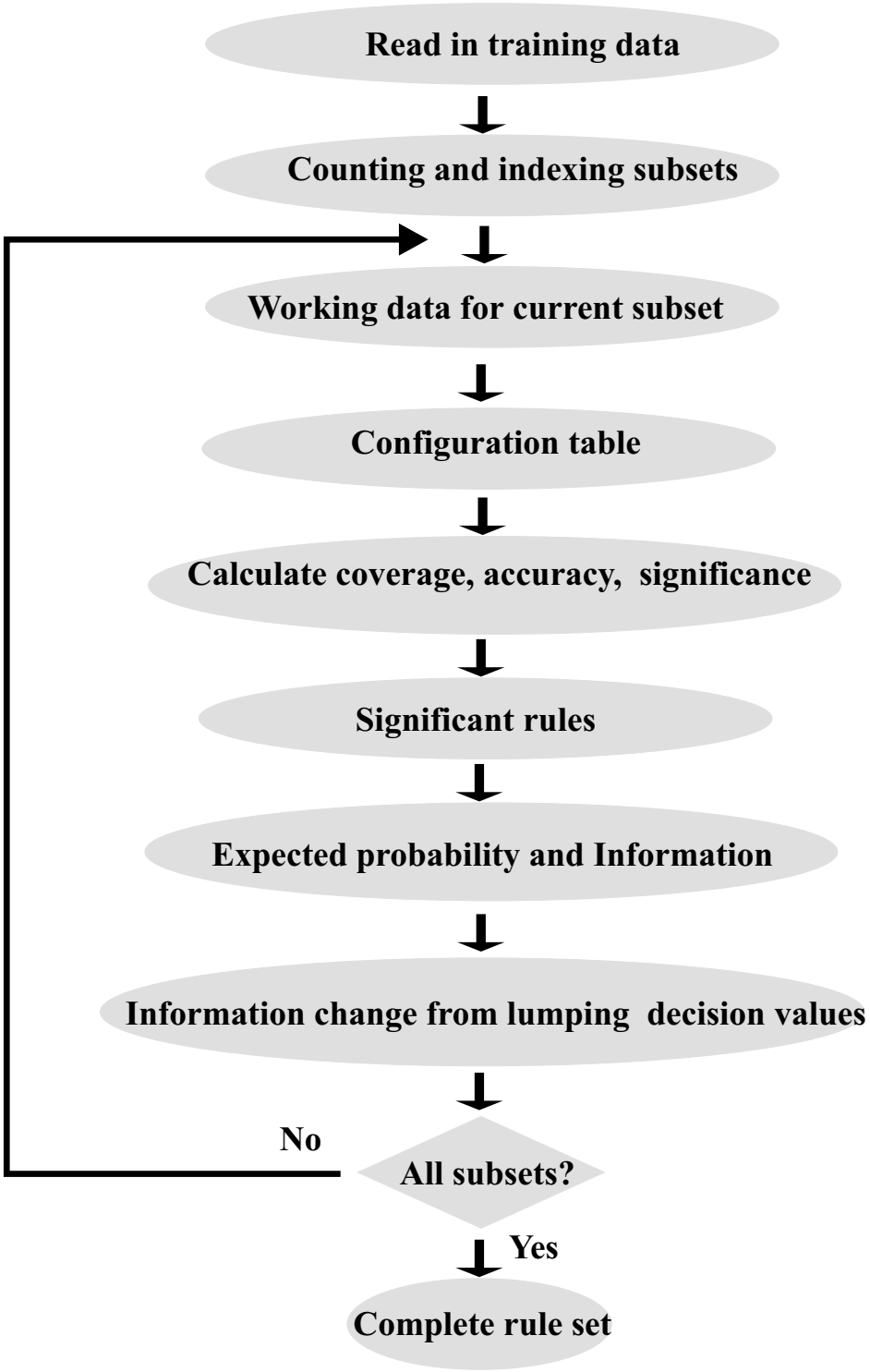

Figure 8: Flow chart of program `ruleind`