I^2AM: a Semi-Automatic System for Data Interpretation in Petroleum Geology

Denis Ferraretti[1], Giacomo Gamberoni[1], and Evelina Lamma[2]

[1] intelliWARE snc, via Borgo dei Leoni 132, 44121 Ferrara, Italy
{denis, giacomo}@i-ware.it
[2] Engineering Department, University of Ferrara, via Saragat 1, 44122 Ferrara, Italy
{evelina.lamma}@unife.it

1 Introduction

The natural complexities of petroleum reservoir systems continue to provide a challenge to geoscientists. In petroleum geology, exploration and production wells are often analysed using image logs and the use of all the available borehole data to completely characterize the reservoir potentials and performance is an important task. The development of reliable interpretation methods is of prime importance regarding the reservoir understanding and data integration is a crucial step in order to create useful description models and to reduce the amount of time necessary for each study. Artificial intelligence, data mining techniques and statistical methods are widely used in reservoir modelling, for instance in prediction of sedimentary facies[3].

The aim of our work was to define and implement a suite of tools for interpretation of image logs and large datasets of subsurface data coming from geological exploration. This led to the development of $\mathbf{I^2AM}$ (Intelligent Image Analysis and Mapping), a semi-automatic system that exploits image processing algorithms and artificial intelligence techniques to analyse and classify borehole data. More in detail, the objectives of the $\mathbf{I^2AM}$ approach are: (1) to automatically extract rock properties information from all the different types of data recorded/measured in the wells, and visual features from image logs in particular; (2) to identify clusters along the wells that have similar characteristics; (3) to predict class distribution over new wells in the same area.

The main benefits of this approach are the ability to manage and use a large amount of subsurface data simultaneously. Moreover, the automatic identification of similar portions of wells by hierarchical clustering saves a lot of time for the geologist (since he analyses only the previously identified clusters). The interpretation time reduces from days to hours and subjectivity errors are avoided. Moreover, chosen clusters are the input for supervised learning methods which learn a classification that can be applied to new wells.

[3] A facies is a body of sedimentary rock distinguished from others by its lithology, geometry, sedimentary structures, proximity to other types of sedimentary rock, and fossil content.

2 The I²AM Approach

With our system, we propose a cascade of techniques, i.e., pattern recognition, clustering and learning classifications algorithms, in order to:

1. first, automatically identify relevant features in image logs, by applying machine vision algorithms;
2. second, cluster several regions of the same well or of different wells into similar groups, by applying hierarchical clustering and choose the set of most significant clusters: this is done by the expert of the domain;
3. finally, feed a machine learning algorithm in order to learn a classifier to be applied to new instances and wells, possibly co-located.

See Figure 1 for the entire workflow.

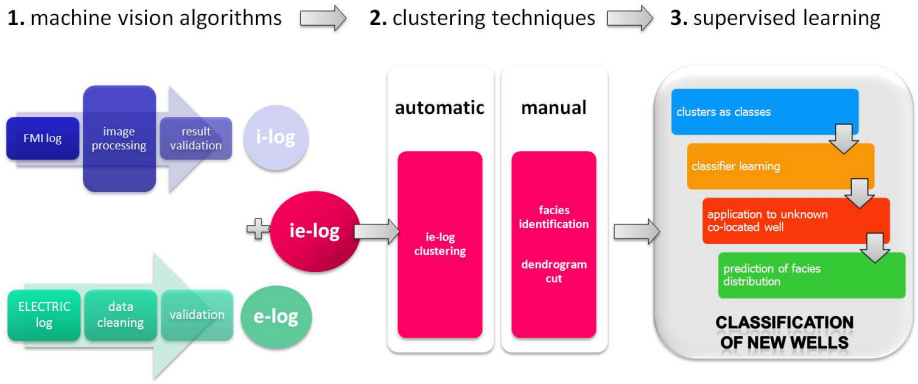


Fig. 1. Workflow of the I²AM system: 1) image logs are analysed using machine vision algorithms and then merged with electrical logs; 2) clustering of the dataset in order to discover hidden data structures; 3) learning and classification of new wells.

In the first step we create a large dataset that includes data from different wells in the same area, this will be the input of following step. Each well is characterized by two types of log: image and electric. In order to use both we need to convert image log observations in numerical dataset. To do this we use machine vision algorithms.

In second step, hierarchical clustering is applied to a set of co-located wells in order to find an hidden data structure. The domain expert chooses the best clustering partition that fits the observed facies distribution. In our application we use hierarchical agglomerative clustering that produces a cluster hierarchy represented in a dendrogram. Using the dendrogram the geologist can choose the most suitable cluster partition.

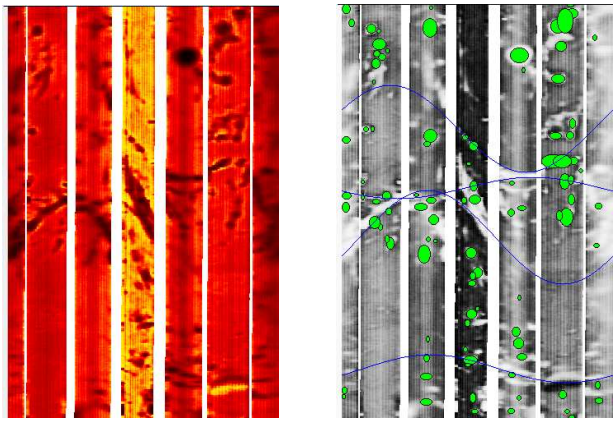
Then in third step, starting from identified clusters, a supervised learning algorithm is used to learn a classifier which can be applied to new wells, in order

to predict the distribution of facies over a new, unknown well in the same area. This task is achieved by learning the model of each cluster from the previous description, to this purpose it is possible to use different supervised techniques.

Following these steps, we obtain a semi-automatic interpretation and prediction method for well logs. This is a semi-automatic approach because a human quality control is needed in order to obtain a meaningful clustering partition in the domain context; but this is also the main advantage: the geologist identifies clusters only once considering all the available data simultaneously and saving time.

2.1 Machine Vision Algorithms

Image logs or FMI[4] logs are digital images acquired by a special logging tool within a borehole [14]. See Figure 2 for an example. FMI logs interpretation is a very complex task, due to the large number of variables and to the huge amount of data to be analysed. Usually, the geologist (domain expert) performs the bedding and fracture analysis by hand, in a tedious and expensive task, and then he tries to identify different classes that group well sections at different depths with similar visual characteristics.



(a) Source FMI image (b) Automatic detection

Fig. 2. Example of FMI image log (a) and automatic extracted features (b): sinusoids (blue curves) and vacuoles (green circles).

The **I²AM** approach for geological image interpretation is based on the detection/measurement of some features for each analysis window (360x100 pixel image), over the entire well.In particular these four features are:

[4] FMI (Fullbore Formation MicroImager) is the name of the tool used to acquire image logs based on resistivity measures within the borehole.

– surfaces (bedding or fracturing that visually correspond to sinusoids);
– vugs/clasts;
– contrast between the previous features and background;
– organization of the texture (homogeneous vs. granular).

In order to classify the features of the images over the entire well, the system analyzes the entire borehole log using an *analysis window* of fixed size. The size of the window is important because it has a direct impact on the resolution of the output/analysis and on the time of analysis of the entire well. The size of this window can be set by the user depending on the type of analysis to be performed.

Sinusoids in the log image can have different geological meanings: bedding or fracture. They do not appear entirely in the FMI, only short parts of them are directly visible.Sinusoids in the log image can have different geological meanings and they are automatically extracted using advanced image processing algorithms developed and tested in in [2] and [3].

To find and count vugs/clasts is important to understand the rock porosity and type of fluid that fills the vacuoles. In the FMI image vacuoles appear as circular or ellipsoidal areas with uniform color, with a high or low contrast with the background. To automatically find and count vugs/clasts the system use algorithms from [4]. The goal is to separate vacuoles from the background and to distinguish them on the basis of some visual features (i.e., area dimension or average color). A trivial count of the vacuoles and sinusoids detected in a zone are fundamental features for the classification of the rock.

The contrast value is significant because it can easily highlight the variation of resistivity in the rock formation. The resistivity variation usually depends on the lithology and the type of rock or type of fluids that fill the pores. This is achieved by using a properly filtered image FFT (Fast Fourier Transform), in order to link to each analyzing window a value that can represent a reliable measure of image contrast.

The internal organization of a rock is an important parameter to understand petrophysics and petrographic characteristics of a rock. The texture organization is highly variable and is an important information for the full interpretation of rock formation, it can be fine-grained to coarse-grained. A grainy FMI image has several small areas (grains) in contrast with the background, and these areas could be highlighted through an edge detection algorithm. The total amount of pixels in the edges of the processed image, is proportional to the texture organization.

Once the system has analysed the entire image log, and the algorithms have extracted the values that represent each feature, these information are summarized in a feature table (a row for each analysis window, a column for each image feature). This table is the final numerical dataset from FMI log and it can be properly merged with other electric logs.

2.2 Clustering Techniques

Cluster analysis is an unsupervised learning method that constitutes a cornerstone of our intelligent data analysis process [10]. It is defined as the task of categorizing objects having several attributes into different classes such that the objects belonging to the same class are similar, and those that are broken down into different classes are not. Intra-connectivity is a measure of the density of connections between the instances of a single cluster. A high intra-connectivity indicates a good clustering arrangement because the instances grouped within the same cluster are highly dependent on each other. Inter-connectivity is a measure of the connectivity between distinct clusters. A low degree of interconnectivity is desirable because it indicates that individual clusters are largely independent of each other. Every instance in the dataset is represented using the same set of attributes.

Generally, clustering algorithms can be categorized into partitioning methods, hierarchical methods, density-based methods, and grid-based methods. In our work we use hierarchical method, it builds the hierarchy starting from the individual elements considered as single clusters, and progressively merges clusters according to a chosen similarity measure defined in features space. Hierarchical clustering techniques use various criteria to decide "locally" at each step which clusters should be joined (or split for divisive approaches). For agglomerative hierarchical techniques, the criterion is typically to merge the "closest" pair of clusters, where "close" is defined by a specified measure of cluster proximity. There are three definitions of the closeness between two clusters: single-link, complete-link and average-link. The single-link similarity between two clusters is the similarity between the two most similar instances, one of which appears in each cluster. Single link is good at handling non-elliptical shapes, but is sensitive to noise and outliers. The complete-link similarity is the similarity between the two most dissimilar instances, one from each cluster. Complete link is less susceptible to noise and outliers, but can break large clusters, and has trouble with convex shapes. The average-link similarity is a compromise between the two. Our application provides best known distance measures: Pearson, Manhattan and Euclidean, and linkage strategies (single, complete and average).

Results of agglomerative algorithms can be represented by dendrograms (see main windows in Figure 3). Advantages of this technique are: 1) it does not require the number of clusters to be known in advance, 2) it computes a complete hierarchy of clusters, 3) good result visualizations are integrated into the methods, 4) a "flat" partition can be derived afterwards (e.g. via a cut through the dendrogram). An excellent survey of clustering techniques can be found in [8].

2.3 Supervised Learning

Inductive machine learning is the process of learning a set of rules from instances (examples in a training set), or more generally speaking, creating a classifier that can be used to generalize from new instances [9].

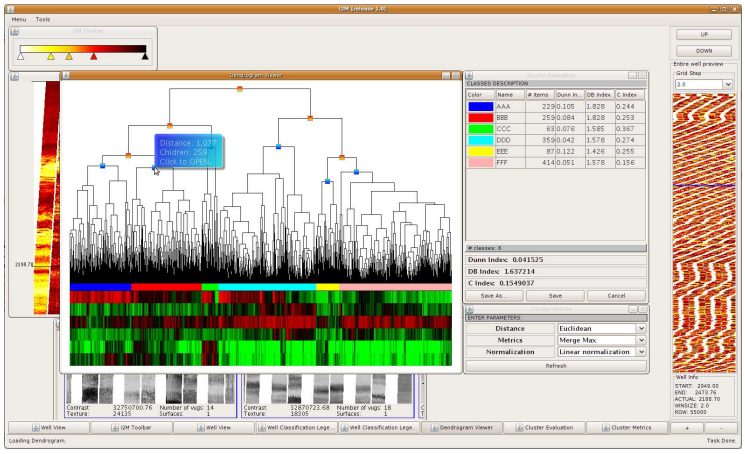


Fig. 3. Clustering process with dendrogram visualization in the I^{2}AM software.

Supervised classification is one of the tasks most frequently carried out by so-called Intelligent Systems. Thus, a large number of techniques have been developed based on artificial intelligence (logical/symbolic techniques), perceptron based techniques and statistics (bayesian networks, instance-based techniques).

In order to find the best classifier for facies distribution prediction in petroleum geology domain, we test several algorithms: decision trees, classification rules and regression methods. These techniques allow the propagation of classes to new wells. We use `J4.8`, `Random Forests`, `PART` [5] and `Rotation Forest` as decision trees induction and classification rules generation algorithms. For regression we use `ClassificationViaRegression` [6] and `Logistic`.

Decision trees represent classification rules in form of a tree, where each node represents a test on an attribute. Depending on the outcome of the test, we must follow the relative branch, and continue until we reach a leaf, that gives a classification of the instance. Decision trees are usually created from examples, using algorithms such as `C4.5` by Quinlan [12]. We use `J4.8` algorithm, which is an implementation of this `C4.5` decision tree learner.

`Random Forests` are a combination of tree predictors such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest [1]. The generalization error for forests converges to a limit as the number of trees in the forest becomes large.

`Rotation Forest` is an algorithm for generating ensembles of classifiers [13]. It consists in splitting the feature set into K subsets, running principal component analysis separately on each subset and then reassembling a new extracted feature set while keeping all the components. The data is transformed linearly into the new features. A decision tree classifier is trained with this data set.

Linear regression can easily be used for classification in domains with numeric attributes. Indeed, we can use any regression technique, whether linear or non-linear, for classification. The trick is to perform a regression for each class, setting

the output equal to one for training instances that belong to the class and zero for those that do not. The result is a linear expression for the class. Then, given a test example of unknown class, calculate the value of each linear expression and choose the one that is largest. This method is sometimes called *multiresponse linear regression*. We use `Logistic`, an implementation of a two-class logistic regression model with a ridge estimator [11]. A complete review of supervised machine learning techniques can be found in [9].

All supervised learning techniques were tested using WEKA, the open source data mining software written in Java [7]. Using several evaluation techniques, detailed in [3], we test classes prediction for 2 wells in a group of 6, and `Logistic` shows better performance than other algorithms in most cases. This result confirm, as expected, that regression methods are suitable for prediction of continuous numeric values.

References

1. L. Breiman, Random Forests, *Machine Learning*, 2001, pp. 5–32.
2. D. Ferraretti, *Data Mining for Petroleum Geology*, PhD Thesis, University of Ferrara, Italy, 2012.
3. D. Ferraretti and G. Gamberoni and E. Lamma, *Unsupervised and supervised learning in cascade for petroleum geology*, Expert Systems with Applications, Elsevier, 2012.
4. D. Ferraretti, Casarotti L., Gamberoni G., E. Lamma, Spot detection in images with noisy background, 16th International Conference on Image Analysis and Processing (ICIAP), Ravenna, Italy, 2011.
5. E. Frank and I. H. Witten, Generating Accurate Rule Sets Without Global Optimization, Proceedings of the Fifteenth International Conference on Machine Learning, 1998, pp. 144–151.
6. E. Frank, Y. Wang, S. Inglis, G. Holmes and I. H. Witten, Using Model Trees for Classification, *Machine Learning*, 1998, pp. 63–76.
7. M. Hall and E. Frank and G. Holmes and B. Pfahringer and P. Reutemann and I.H. Witten, *The WEKA data mining software: an update*, ACM SIGKDD Explorations Newsletter, Vol. 11, No. 1, pp. 10–18, ACM, 2009.
8. A.K. Jain and M.N. Murty and P.J. Flynn, *Data clustering: a review*, ACM computing surveys (CSUR), Vol. 31, No. 3, pp. 264–323, ACM, 1999.
9. S.B. Kotsiantis and I.D. Zaharakis and P.E. Pintelas, *Supervised machine learning: A review of classification techniques*, Frontiers in Artificial Intelligence and Applications, Vol. 160, IOS Press, 2007.
10. S. B. Kotsiantis and P.E. Pintelas, *Recent advances in clustering: A brief survey*, WSEAS Transactions on Information Science and Applications, Vol. 1, No. 1, pp. 73–81, Citeseer, 2004.
11. S. Le Cessie, J. C. Van Houwelingen, Ridge Estimators in Logistic Regression, *Applied Statistics*, 1992.
12. J. R. Quinlan, Induction on Decision Trees, *Machine Learning*, 1986, pp. 81–106.
13. J. J. Rodriguez, L. I. Kuncheva, C. J. Alonso, Rotation Forest: A New Classifier Ensemble Method, *IEEE Trans. Pattern Anal. Mach. Intell.*, 2006, pp. 1619–1630.
14. O. Serra and L. Serra, *Well Logging, Data Acquisition and Applications*, SerraLog, 2004.