di4g: Uno strumento di *clustering* per l'analisi integrata di dati geologici

Alice Piva[1], Giacomo Gamberoni[1], Denis Ferraretti[1], Evelina Lamma[2]

[1] intelliWARE snc, via J.F.Kennedy 15, 44122 Ferrara, Italia
{denis, giacomo}@i-ware.it, alice88.piva@gmail.com
[2] Dipartimento di Ingegneria, Università di Ferrara, via Saragat 1, 44122 Ferrara, Italia
{evelina.lamma}@unife.it

1 Introduzione

di4g (data integrator for geology) è uno strumento sviluppato per l'analisi di dati geologici, in particolare per la geologia degli idrocarburi. Questa disciplina si occupa di cercare e valutare gli elementi fondamentali nella formazione di un giacimento di idrocarburi in un bacino sedimentario. Tipicamente, attraverso sonde calate nei pozzi esplorativi, si ottengono numerosi dati eterogenei, costituiti da *log* elettrici e rappresentati sotto forma di curve e *log* immagini (detti FMI) significative della conformazione delle pareti dei pozzi. Da queste immagini si possono ricavare informazioni riguardanti la tessitura delle rocce, il tipo di porosità, la presenza di fratture (rappresentate da sinusoidi). L'esperto geologo analizza questi *log* visivamente, per identificare le varie caratteristiche presenti all'interno delle immagini. Questa è però un'analisi complessa e soggettiva nell'interpretazione che richiede, inoltre, un elevato tempo di esecuzione. Per questo, è stato sviluppato I^2AM (Intelligent Image Analysis and Mapping), un software per l'interpretazione semiautomatica delle immagini provenienti dai pozzi petroliferi (1) e per individuare le caratteristiche visuali presenti (6, 8). Poiché tutti i dati disponibili sono eterogenei e non tutti considerati da I^2AM, per consentire l'allineamento e la fusione di diversi *dataset* è stato sviluppato di4g. di4g fonde la tabella delle caratteristiche prodotta da I^2AM con i *log* elettrici disponibili e consente anche di eseguire l'analisi integrata di dati provenienti da pozzi diversi.
Per fornire una prima classificazione sui dati in ingresso, di4g applica una tecnica di *clustering* individuando zone del pozzo simili e raggruppandole in *cluster* con un algoritmo di *clustering*. Ciò costituisce una prima ipotesi di divisione delle parti dei pozzi in zone simili, in ciascuna delle quali sono presenti diverse tipologie di roccia. Tramite l'utilizzo del software di4g il geologo può quindi eseguire un'interpretazione molto più veloce e interattiva dei dati ricavati dalle esplorazioni, grazie anche alle rappresentazioni grafiche di uscita dello strumento aggiornate in tempo reale. Sono stati sperimentati diversi approcci al fine di individuare -e in alcune situazioni prevedere- le classificazioni per determinati campi di pozzi (4, 5, 7, 9, 10), ma l'approccio non supervisionato con il coinvolgimento diretto del geologo si è dimostrato il più efficace (2). di4g è stato sviluppato in *partnership* tra due aziende, una software e una di consulenza geologica, e un Dipartimento Universitario di Ingegneria. La versione

di di4g descritta è stata sviluppata nel corso di un tirocinio e tesi di laurea magistrale svolti da parte del primo autore.

2 L'approccio di di4g

L'obiettivo del progetto è stato quello di sviluppare uno strumento per l'analisi integrata di dati da esplorazioni petrolifere, orientato allo studio di diversi pozzi e di tipi di dati eterogenei (3). In genere, in fase di analisi di un terreno, in un campo esplorativo si eseguono più trivellazioni di pozzi. I dati devono essere analizzati e confrontati simultaneamente per identificare quali tra questi pozzi hanno le caratteristiche migliori per essere un buon giacimento.
Il flusso di lavoro di di4g è suddiviso in tre fasi principali (si veda Figura 1):

- nella prima fase avviene l'**importazione** dei file. L'applicazione consente di importare più file contenenti dati riferiti ad ogni pozzo e tramite un algoritmo di allineamento costruisce il *dataset* utilizzato nel successivo processo di *clustering*. Altra funzionalità fondamentale in questa fase è la possibilità di importare file contenenti dati riferiti a più pozzi; in questo modo è possibile analizzare contemporaneamente le caratteristiche di un intero campo esplorativo;
- nella seconda fase avviene il processo di **clustering**. In questa fase si normalizzano i dati, e la normalizzazione è solitamente applicata sull'intero *dataset*. Avendo inoltre la possibilità di importare dati riferiti a più pozzi, il sistema è stato reso in grado di calcolare anche una normalizzazione riferita a ogni singolo pozzo. In questa fase è stato anche ottimizzato il processo di *clustering*.
- nella terza fase avviene l'**esportazione** del *dataset* risultante dal processo di *clustering*. In questa fase è importante per il geologo poter costruire diversi tipi di grafici a partire dal *dataset* ottenuto dal *clustering*. I grafici aiutano il geologo nel processo di identificazione dei tipi di rocce del pozzo esplorato. Sono stati anche integrati direttamente nell'applicazione di4g moduli di visualizzazione dei grafici e una tabella contenente le statistiche per ogni singolo *cluster* creato.

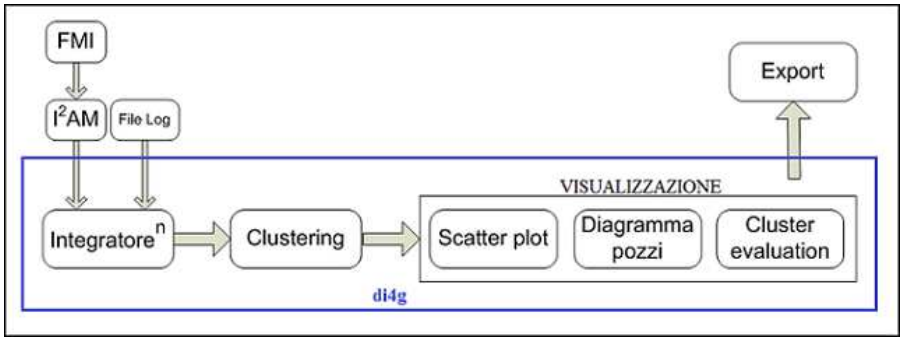

Fig. 1. Workflow di di4g

2.1 Importazione e integrazione dei dati

L'applicazione fornisce un'interfaccia che consente l'importazione di file provenienti da tecniche e strumenti diversi, per numerosi pozzi. Il file d'ingresso dovrà essere una tabella numerica in formato testo con tante righe quante sono le profondità analizzate e tante colonne (separate da un delimitatore o lunghezza fissa) quante sono le variabili prese in esame riguardanti ad esempio la densità e la porosità delle rocce; dovrà inoltre contenere gli identificativi dei diversi pozzi. Un'analisi globale dell'insieme di pozzi, fornisce una classificazione unificata e quindi una rappresentazione coerente delle tipologie di rocce (questo è particolarmente utile solo per pozzi geograficamente vicini).
Come detto in precedenza i dati possono essere generati utilizzando strumenti differenti e avere così anche risoluzioni differenti; il sistema fornisce all'utente la possibilità di importare molti formati di testo, potendo scegliere separatore, numero di righe d'intestazione, e altre caratteristiche.
Inoltre è necessario in questa prima fase eseguire un allineamento dei diversi file importati. L'allineamento avviene analizzando ogni profondità della matrice, ricavata dai dati del primo file, con lo scopo di trovare per ognuna di esse tutte le righe della seconda matrice la cui profondità è "vicina" a quella analizzata. Così facendo l'elenco delle profondità del primo file fa da "guida" per tutti i file. Durante questa fase ci si può inoltre trovare nella situazione in cui questa debba avvenire tra dati di un pozzo che però presentano profondità iniziali e finali che non coincidono: di4g effettua quindi una 'pulizia' delle profondità iniziali e finali della matrice da allineare e si ottiene in questo modo un allineamento molto più uniforme e coerente delle profondità iniziali e finali di ogni pozzo.

2.2 La fase di clustering

Nella seconda fase, di4g esegue il processo di *clustering*. Con il termine *clustering* si intende un insieme di tecniche volte a individuare e raggruppare gli elementi omogenei in un insieme di dati; in particolare, il *clustering* ha l'obiettivo di trovare in un *dataset* dei gruppi che siano il più differenti possibili gli uni dagli altri, ma allo stesso tempo con membri di uno stesso gruppo il più possibile simili tra loro.
Come tecnica di *clustering*, di4g utilizza il *clustering* gerarchico agglomerativo non supervisionato (11), cioè un approccio dal basso verso l'alto che parte inserendo ogni elemento in un *cluster* differente, iterativamente calcola la distanza tra i diversi *cluster* e fonde i *cluster* che si trovano a distanza minore, ossia quelli più simili; il procedimento è iterato fino ad ottenere un unico *cluster*. Lo strumento realizzato consente di scegliere la distanza da utilizzare per misurare la similarità tra due elementi (sono disponibili la distanza di Manhattan, Euclidea e Pearson); inoltre permette di scegliere fra tre diverse metriche (Merge Max, Merge Min, Merge Avg) da utilizzare per selezionare la coppia di *cluster* da fondere.
La normalizzazione è importante per rendere il risultato indipendente dalle unità di misura delle variabili, facendo in modo che tutte le variabili contribuiscano in ugual misura; in questo modo si esegue il calcolo delle distanze su variabili che sono con-

frontabili. Oltre a una classica normalizzazione sull'intero *dataset* è stata realizzata anche la possibilità di eseguire una normalizzazione su ogni singolo pozzo. La normalizzazione sull'intero *dataset* è consigliata quando si è in presenza di dati uniformi (ad esempio se sono stati usati gli stessi strumenti e parametri per la rilevazione); la normalizzazione su ogni singolo pozzo è invece la modalità appropriata nel caso in cui sono stati definiti parametri e tarature differenti degli strumenti utilizzati per l'estrazione dei dati riferiti ai diversi pozzi: così facendo i dati vengono allineati e portati nello stesso *range* correggendo le diverse tarature utilizzate.
In di4g il processo di *clustering* è realizzato tramite la costruzione di una matrice delle distanze in cui si cerca il minimo. Tale matrice può occupare uno spazio considerevole in memoria (ordine $O(N^2)$ dove N è il numero di foglie presenti nel *dataset*) se siamo in presenza di *dataset* di grandi dimensioni e la ricerca del minimo all'interno di tale matrice può diventare un processo oneroso. Per questo motivo di4g ottimizza la ricerca del minimo all'interno di questa matrice tramite alcuni vettori ausiliari. Una quota importante del lavoro si è concentrata nell'analizzare e ridurre i tempi del processo di *clustering*; in particolare l'ottimizzazione del popolamento di questi vettori ha ridotto drasticamente i tempi di esecuzione.
Il risultato del *clustering* gerarchico è un diagramma ad albero chiamato dendrogramma sul quale è possibile generare un taglio per ottenere una configurazione di *cluster*. La finestra di uscita prodotta da di4g consente di eseguire due diversi tipi di taglio: il classico taglio orizzontale e il taglio obliquo (rappresentato in Figura 2). Il taglio obliquo permette di tagliare l'albero ad altezze diverse, semplicemente selezionando i nodi che si desiderano "aprire". In questo modo è possibile espandere in profondità una sola parte dell'albero, e il geologo ha quindi la flessibilità di scegliere la configurazione di *cluster* per lui più rappresentativa.

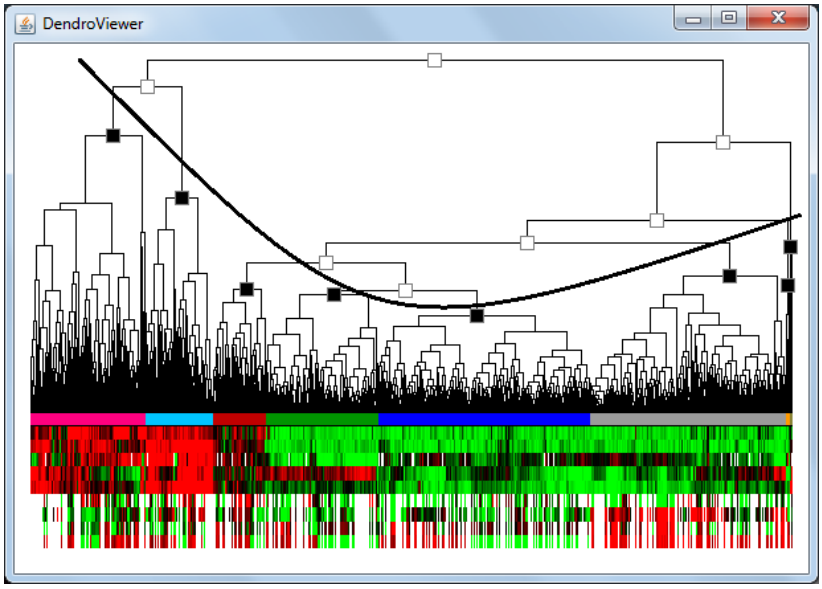

Fig. 2. Dendrogramma con esempio di taglio obliquo (nodi in nero)

2.3 Visualizzazione interattiva

La terza fase di di4g consente la visualizzazione di vari tipi di grafici interattivi aggiornabili simultaneamente nel momento in cui si cambia la configurazione di *cluster* agendo sul dendrogramma oppure se è avviata una nuova procedura di *clustering*.
Uno dei grafici visualizzabili è il diagramma pozzo: rappresentazione del pozzo in verticale con ogni profondità colorata secondo il *cluster* di appartenenza (Figura 3). Di questo diagramma si possono visualizzare diverse configurazioni: su un solo pozzo oppure su pozzi multipli, in questo caso possono essere rappresentati allineati a partire dall'alto oppure scalati secondo la loro profondità reale. Per il geologo il diagramma pozzo è fondamentale in quanto gli consente di individuare le caratteristiche del pozzo alle diverse profondità, in particolare in caso di diagramma su pozzi multipli, il geologo può effettuare un confronto simultaneo e individuare più rapidamente le varie caratteristiche dei diversi pozzi.

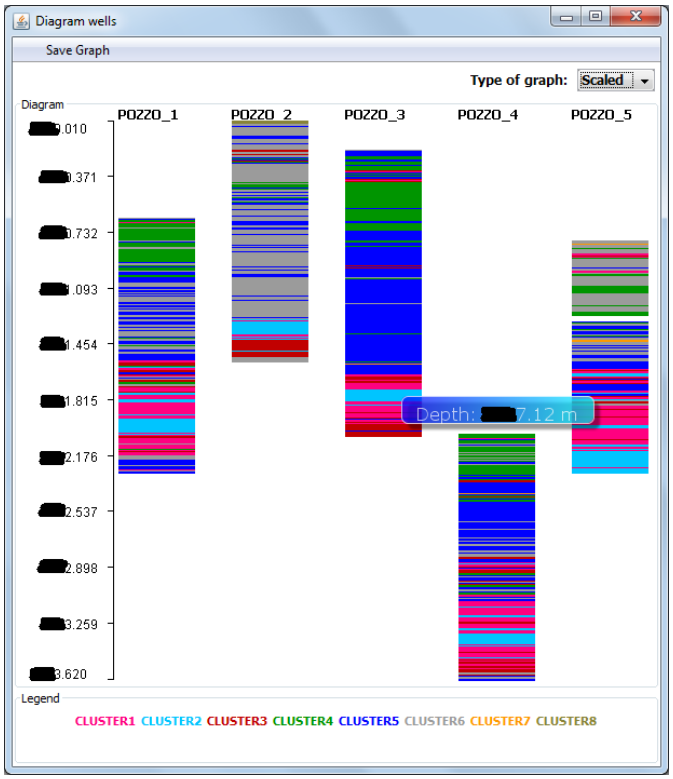

Fig. 3. Diagramma pozzo

Con il grafico di correlazione (Figura 4), di4g rappresenta in uno spazio cartesiano il grado di dipendenza tra due variabili (si possono considerare tutti i cluster contemporaneamente o uno specifico *cluster*).

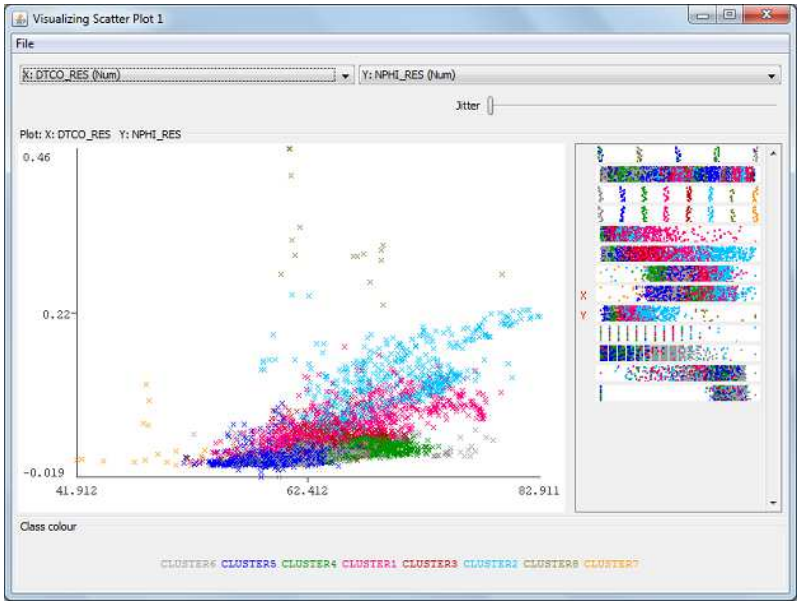

Fig. 4. Scatter plot

Un'ultima funzionalità permette di visualizzare una tabella di 'Cluster Evaluation' riassuntiva delle statistiche dei *cluster*; per ognuno di essi il geologo può analizzare ogni variabile presa in esame. Per semplificare ulteriormente l'interpretazione effettuata dal geologo dalla tabella di 'Cluster Evaluation' per ogni singola variabile è possibile visualizzare il box-plot che consente di rappresentare la distribuzione di una determinata variabile all'interno dei diversi *cluster* e confrontarle tra loro; è inoltre possibile visualizzare anche la distribuzione di una determinata variabile in un *cluster* all'interno dell'istogramma.
Come detto in precedenza tutti i grafici descritti si rigenerano in seguito al cambiamento del taglio sul dendrogramma, in questo modo, l'analisi svolta dal geologo può essere maggiormente interattiva e molto più veloce.

3 Conclusioni

di4g è uno strumento che permette l'importazione di file che contengono grandi quantità di dati e la loro integrazione con dati visuali, per eseguire poi il processo di *clustering*. Questo strumento presenta diversi vantaggi: in primo luogo può essere utilizzato anche da utenti non esperti grazie ad un'interfaccia semplice e intuitiva. Inoltre, avendo la possibilità di importare file contenenti dati riferiti a più pozzi, permette di analizzare le caratteristiche dei diversi pozzi contemporaneamente, effettuando un confronto simultaneo e riducendo così i tempi di analisi. Un ulteriore vantaggio è dato dalla possibilità di visualizzare grafici che consentono di semplificare e velocizzare il lavoro del geologo analista. Ogni tipo di grafico realizzato porta dei vantaggi in fase

di analisi, infatti, è possibile verificare i legami tra due variabili analizzate, eseguire un confronto istantaneo sulle distribuzioni dei diversi pozzi, verificare la distribuzione dei *cluster* lungo la profondità di ogni singolo pozzo. Il loro aggiornamento simultaneo consente al geologo di avere sott'occhio costantemente le modifiche derivanti da una nuova configurazione di *cluster*. Questo consente un'interpretazione molto più veloce di ogni singola zona del pozzo e di verificare così quale tra le diverse configurazioni rappresenta al meglio i dati importati. Infine di4g è nato in ambito geologico, ma a nostro parere può essere utilizzato in diversi ambiti, ovvero in tutti quelli in cui si ha la necessità di eseguire un'analisi su dei dati in ingresso.

Riferimenti

1. **Denis Ferraretti; Giacomo Gamberoni; Evelina Lamma.** I2AM: a Semi-Automatic System for Data Interpretation in Petroleum Geology. PAI 2012, CEUR WS Vol.860.
2. **Denis Ferraretti.** *Data Mining for Petroleum Geology.* Tesi di Dottorato di Ricerca in Scienze dell'Ingegneria. Università degli Studi di Ferrara, 2012.
3. **Alice Piva.** *Sviluppo e ottimizzazione di uno strumento di clustering per la geologia.* Tesi di Laurea Magistrale in Ingegneria Informatica e dell'Automazione. Università degli Studi di Ferrara, 2012.
4. **Denis Ferraretti; Giacomo Gamberoni; Evelina Lamma.** Unsupervised and supervised learning in cascade for petroleum geology. Expert Syst. Appl. 2012, Volume 39, Issue 10, pp. 9504-9514. http://dx.doi.org/10.1016/j.eswa.2012.02.104
5. **Denis Ferraretti; Evelina Lamma; Giacomo Gamberoni; Michele Febo.** Clustering and Classification Techniques for Blind Predictions of Reservoir Facies. AI*IA 2011, LNAI 6934, pp. 348-359, Springer.
6. **Denis Ferraretti; Luca Casarotti; Giacomo Gamberoni; Evelina Lamma.** Spot Detection in Images with Noisy Background. ICIAP 2011, LNCS 6978, pp. 575-584, Springer.
7. **Denis Ferraretti; Evelina Lamma; Giacomo Gamberoni; Michele Febo; Raffaele Di Cuia.** Integrating Clustering and Classification Techniques: A Case Study for Reservoir Facies Prediction. ISMIS 2011, Emerging Intelligent Technologies in Industry, Studies in Computationl Intelligence 369, pp. 21-34, Springer.
8. **Denis Ferraretti; Luca Tagliavini; Raffaele Di Cuia; Mariachiara Puviani; Evelina Lamma; Sergio Storari.** Use Of Artificial Intelligence Techniques To The Interpretation Of Subsurface Log Images. Intelligenza Artificiale 2010, AI*IA, pp. 27-35, IOS Press.
9. **Denis Ferraretti; Giacomo Gamberoni; Evelina Lamma.** Automatic Cluster Selection Using Index Driven Search Strategy. *AI*IA 2009: Emergent Perspectives in Artificial Intelligence.* LNAI 5883, pp 172-181,Springer.
10. **Denis Ferraretti; Giacomo Gamberoni; Evelina Lamma; Raffaele Di Cuia; Chiara Turolla.** An AI Tool for the Petroleum Industry Based on Image Analysis and Hierarchical Clustering. IDEAL 2009, LNCS 5788, pp. 276-283, Springer.
11. **Stephen C. Johnson.** Hierarchical clustering schemes. *Psychometrika.* 1967, Volume 32, Issue 3 , pp 241-254.