summer s.r.l.
SISTEMI FINANZIARI OTTIMIZZATI

Reg. Soc. Tib. Brescia n. 40188 – Codice Fiscale: 07073330156 - Partita IVA: 03051110173 – C.C.I.A.A. 314779
Sede: Viale Italia, 1 – 25126 Brescia – Fax 030 5109179 – Tel. 030 2400330 (5 linee con r.a.)
Sito web: www.summer.it – E-mail: summer@summer.it

e-GRID

Sistema per l'elaborazione automatica

di griglie statistiche di previsione

Riferimento: 8 marzo 2007

summer s.r.l.
SISTEMI FINANZIARI OTTIMIZZATI

Reg. Soc. Trib. Brescia n. 40188 – Codice Fiscale: 07073330156 - Partita IVA: 03051110173 – C.C.I.A.A. 314779
Sede: Viale Italia, 1 – 25126 Brescia – Fax 030 5109179 – Tel. 030 2400330 (5 linee con r.a.)
Sito web: www.summer.it – E-mail: summer@summer.it

Indice:

Reg. Soc. Trib. Brescia n. 40188 – Codice Fiscale: 07073330156 - Partita IVA: 03051110173 – C.C.I.A.A. 314779
Sede: Viale Italia, 1 – 25126 Brescia – Fax 030 5109179 – Tel. 030 2400330 (5 linee con r.a.)
Sito web: www.summer.it – E-mail: summer@summer.it

1 OBIETTIVI

Il presente documento descrive le principali caratteristiche e modalità di utilizzo del Sistema Informativo "*e*-GRID" per l'elaborazione automatica di griglie statistiche di previsione, fornendo il dettaglio tecnico dei processi di calcolo supportati.

2 DESCRIZIONE DEL SISTEMA "*e*-GRID"

Il Sistema "*e*-GRID" è finalizzato all'elaborazione di griglie statistiche mediante cui effettuare previsioni circa l'andamento di specifiche funzioni, sulla base dell'analisi di set di osservazioni storiche e attraverso l'applicazione di tecniche di "data mining".

Il Sistema si configura come un "laboratorio elettronico", nell'ambito del quale vengono preparati ed eseguiti specifici esperimenti finalizzati alla determinazione di "funzioni soluzione" tendenzialmente ottimali.

Al fine di assistere e guidare l'analista nel corso delle attività sperimentali, l'applicativo supporta i seguenti moduli:

gestione delle "funzioni";
alimentazione del "data source";
normalizzazione del "data source";
generazione dei "datamart";
interpretazione dei "datamart";
esecuzione degli "esperimenti";
consultazione dei "risultati".

Reg. Soc. Trib. Brescia n. 40188 – Codice Fiscale: 07073330156 - Partita IVA: 03051110173 – C.C.I.A.A. 314779
Sede: Viale Italia, 1 – 25126 Brescia – Fax 030 5109179 – Tel. 030 2400330 (5 linee con r.a.)
Sito web: www.summer.it – E-mail: summer@summer.it

2.1 GESTIONE DELLE "FUNZIONI"

La prima fase dell'attività sperimentale è rappresentata dall'identificazione della "funzione" che s'intende stimare.

Una funzione viene identificata mediante un "nome" ed una "descrizione". Il modulo gestionale consente all'Analista di censire un numero illimitato di funzioni.

Da un punto di vista logico, selezionare una funzione significa circoscrivere un contesto nell'ambito del quale effettuare le sperimentazioni. Alla funzione risulteranno infatti associate le varie entità coinvolte nell'analisi: "data mart", "interpretazioni", "esperimenti" e "risultati".

Il fatto di poter gestire più contesti testimonia la possibilità di impiegare i tools supportati dal Sistema al fine di risolvere problemi di classificazione all'interno di domini differenziati.

Esempi di funzioni sono:

la Probability of Default (PD) e la Loss Gigen Default (LGD), nell'ambito delle scomposizioni in pool inerenti la normativa "Basilea 2";

il Tempo Medio di Recupero (TMR), per quanto attiene le problematiche afferenti le Stime di Perdita in ambito IAS.

Reg. Soc. Trib. Brescia n. 40188 – Codice Fiscale: 07073330156 - Partita IVA: 03051110173 – C.C.I.A.A. 314779
Sede: Viale Italia, 1 – 25126 Brescia – Fax 030 5109179 – Tel. 030 2400330 (5 linee con r.a.)
Sito web: www.summer.it – E-mail: summer@summer.it

2.2 ALIMENTAZIONE DEL "DATA SOURCE"

Dopo aver definito il contesto di sperimentazione, l'Analista è chiamato ad alimentare il "data source", vale a dire la base di osservazioni storiche a cui intende attingere per generare i "data mart" presso cui saranno effettuati gli "esperimenti".

Da un punto di vista tecnico, il "data source" è un file sequenziale il cui tracciato si articola in cento colonne aventi formato alfanumerico e lunghezza trenta, ciascuna delle quali codifica una variabile di ingresso. Ogni riga del "data source" rappresenta un'osservazione storica. La genericità di tale struttura svincola il Sistema dalla particolare forma delle interfacce che possono alimentarlo, garantendone l'apertura verso i Sistemi di Origine.

Il modulo di alimentazione consente all'Analista di:

connettersi ad un Sistema Origine remoto, specificando l'"indirizzo IP" ed indicando la "login" e la "password" di accesso;
selezionare un flusso (strutturalmente compatibile con il "data source") presso il file system remoto e trasferirlo verso il Sistema di destinazione;
popolare il "data source" attingendo ad uno o più flussi importati.

Si precisa che il "data source" è unico, pertanto ogni nuova "popolazione" dello stesso cancella i dati precedentemente caricati.

Reg. Soc. Trib. Brescia n. 40188 – Codice Fiscale: 07073330156 - Partita IVA: 03051110173 – C.C.I.A.A. 314779
Sede: Viale Italia, 1 – 25126 Brescia – Fax 030 5109179 – Tel. 030 2400330 (5 linee con r.a.)
Sito web: www.summer.it – E-mail: summer@summer.it

2.3 NORMALIZZAZIONE DEL "DATA SOURCE"

Le informazioni caricate nel "data source" possono essere soggette ad incompletezze dovute a discontinuità nei valori assunti da talune colonne. L'unità di "normalizzazione" consente di definire le logiche in base alle quali rimuovere tali discontinuità preliminarmente alla generazione dei "data mart".

Il processo di "normalizzazione" avviene per "completamento" dei valori mancanti, mediante tecniche di "interpolazione" applicate alla colonna obiettivo (oggetto di normalizzazione) nell'ambito del set di righe aggregate rispetto all'invarianza di una colonna cardine ed in base all'andamento dei valori di una colonna indipendente, oppure per "cancellazione" immediata delle righe che esibiscono discontinuità.

L'Analista ha facoltà di censire regole di completamento o cancellazione specificando:

per le regole di completamento:
- la colonna obiettivo;
- il valore identificante le discontinuità;
- la colonna indipendente;
- la colonna cardine;

per le regole di cancellazione:
- la colonna obiettivo;
- il valore identificante le discontinuità e caratterizzante le righe che devono essere eliminate.

Una volta terminato il censimento delle regole, l'Analista avvia il processo di normalizzazione. Si precisa che tale processo altera il contenuto del "data source" il quale, se necessario, potrà essere ripristinato solamente mediante una nuova importazione dei flussi.

Reg. Soc. Trib. Brescia n. 40188 – Codice Fiscale: 07073330156 - Partita IVA: 03051110173 – C.C.I.A.A. 314779
Sede: Viale Italia, 1 – 25126 Brescia – Fax 030 5109179 – Tel. 030 2400330 (5 linee con r.a.)
Sito web: www.summer.it – E-mail: summer@summer.it

2.4 GENERAZIONE DEI "DATA MART"

Il "data source" esibisce, per definizione, una struttura poco articolata, necessaria a garantirne la semplicità di alimentazione. Per questa ragione, i dati ivi contenuti sono scarsamente manipolabili.

Ne discende l'esigenza di trasformare il "data source" in uno o più archivi statistici, i cui campi, detti "dimensioni", siano caratterizzati da range maggiormente specializzati e costituiscano aggregazioni "intelligenti" (vale a dire utili ai fini del perseguimento dell'obiettivo) delle colonne del "data source" stesso. Tali archivi, contraddistinti da un alto numero di record ("fatti") aventi tracciato ridotto, vengono detti "data mart".

La trasformazione del "data source" in "data mart" avviene mediante gruppi di regole SQL aggregate in specifiche "procedure di generazione". L'Analista ha facoltà di censire un numero illimitato di procedure specificando la lista delle "regole di trasformazione" che le compongono. Una "regola di trasformazione" definisce le modalità di creazione e alimentazione delle singole "dimensioni" del "data mart" in funzione dei valori contenuti in talune colonne del "data source". Una regola è caratterizzata dai seguenti attributi:

codice della nuova dimensione;
descrizione mnemonica della dimensione;
regola SQL di alimentazione della dimensione.

La regola SQL aggrega, utilizzando gli operatori matematici o di manipolazione testuale supportati dal linguaggio SQL, i valori rivenienti da talune colonne del "data source" in un'unica dimensione.

Il Sistema consente di "testare" una regola prima di applicarla, fornendo l'immagine dei primi cento risultati della "dimensione" che verrebbe generata.

La creazione di un "data mart" avviene selezionando ed eseguendo una "procedura di generazione". Il "data mart" viene contestualmente identificato mediante un codice ed una descrizione.

Reg. Soc. Trib. Brescia n. 40188 – Codice Fiscale: 07073330156 - Partita IVA: 03051110173 – C.C.I.A.A. 314779
Sede: Viale Italia, 1 – 25126 Brescia – Fax 030 5109179 – Tel. 030 2400330 (5 linee con r.a.)
Sito web: www.summer.it – E-mail: summer@summer.it

2.5 INTERPRETAZIONE DEI "DATA MART"

Poiché il Sistema "*e*-GRID" ha l'obiettivo di supportare l'Analista nell'ambito della risoluzione di problemi di classificazione ai fini della stima di talune funzioni, particolare rilevanza riveste l'attività di aggregazione dei valori inclusi nei domini delle variabili rispetto alle quali la funzione indagata sarà potenzialmente espressa.

Tale attività, detta di "interpretazione", si traduce nella definizione, per un sottoinsieme delle "dimensioni" del "data mart" esaminato, di una famiglia di "partizioni". Una "partizione" è rappresentata da un set esaustivo di "unioni di intervalli" disgiunte e incluse nel dominio della "dimensione". In tal senso, una "partizione" aggrega i valori assumibili da una "dimensione" in base a logiche dipendenti dal particolare contesto applicativo. Come vedremo, sarà il Sistema a selezionare tra quelle ipotizzate dall'Analista, le "partizioni" maggiormente performanti ai fini della stima della funzione obiettivo,.

L'Analista ha facoltà di:

indicare il "data mart" che intende "interpretare";
censire, per tale "data mart", nuove "interpretazioni";
nell'ambito di una "interpretazione", definire le "dimensioni" del "data mart" che dovranno essere utilizzate dagli "esperimenti" facenti riferimento all'"interpretazione" (le rimanenti verranno trascurate, dando per scontato che la funzione obiettivo sia costante rispetto ad esse);
per ogni "dimensione" censita, gestire le "partizioni" disponibili;
nell'ambito di una "partizione", caricare le "unioni" che la compongono;
per ogni "unione", gestire gli "intervalli" che la definiscono;
creare nuovi "intervalli" di natura propria (segmenti della retta reale) o impropria (singoletti numerici/alfanumerici), utilizzabili in fase di composizione delle "unioni".

Al fine di facilitare le attività di configurazione, l'Analista dispone di una funzione di copiatura di impianti esistenti a livello di singola "interpretazione", "dimensione", "partizione" o "unione".

Reg. Soc. Trib. Brescia n. 40188 – Codice Fiscale: 07073330156 - Partita IVA: 03051110173 – C.C.I.A.A. 314779
Sede: Viale Italia, 1 – 25126 Brescia – Fax 030 5109179 – Tel. 030 2400330 (5 linee con r.a.)
Sito web: www.summer.it – E-mail: summer@summer.it

2.6 ESECUZIONE DEGLI ESPERIMENTI

Generato il "data mart" e specificata l'"interpretazione" di riferimento, l'Analista è nelle condizioni di eseguire un "esperimento", vale a dire di invocare il Sistema, il quale, mediante uno specifico algoritmo di "data mining" le cui dinamiche verranno approfondite nel prosieguo del documento, determina una o più funzioni adatte a stimare la funzione obiettivo indagata. Ogni funzione viene rappresentata mediante un "albero di classificazione" i cui nodi corrispondono a sottoinsiemi del "data mart", detti "classi".

Operativamente, l'Analista ha la possibilità di inserire e configurare un nuovo "esperimento" specificando:

la descrizione mnemonica dell'attività sperimentale;
il "data mart" nell'ambito del quale eseguirla;
l'"interpretazione" di riferimento;
la "dimensione" del "data mart" che svolge il ruolo della "misura", vale a dire dell'output della funzione indagata;
la "partizione" della "misura" che dovrà essere utilizzata come "scala" per esprimere i valori della funzione e, per ogni elemento di tale "partizione", il relativo "rappresentante" numerico;
la metodologia di aggregazione di tali "rappresentanti" (media o mediana) nell'ambito della singola "classe";
la modalità di inclusione dei "fatti" nel set dedicato al training (casuale, puntuale, ecc.) e gli eventuali relativi parametri di controllo;
la dimensione rispetto alla quale effettuare eventuali "amplificazioni" dei "fatti" ed il criterio da utilizzarsi (lineare, Fibonacci o esponenziale);
la "profondità massima" degli "alberi di classificazione" generabili;
il numero minimo di iterazioni previste per l'"esperimento";
il criterio di arresto dell'algoritmo (soglia entropia di test, soglia entropia di training o numero di epoche).

Gli "esperimenti" configurati possono essere eseguiti. L'esecuzione di un "esperimento" comporta l'avvio dell'algoritmo di "data mining" un numero di volte pari alle iterazioni minime specificate, sebbene sia possibile ripetere l'esecuzione a piacimento. Ogni iterazione comporta la generazione di un "risultato", vale a dire di un "albero di classificazione" rappresentabile in formato XML. Non è ammessa alcuna variazione nella configurazione qualora anche un solo "risultato" insista sull'"esperimento". Gli esiti elaborati possono in ogni caso essere cancellati.

Reg. Soc. Trib. Brescia n. 40188 – Codice Fiscale: 07073330156 - Partita IVA: 03051110173 – C.C.I.A.A. 314779
Sede: Viale Italia, 1 – 25126 Brescia – Fax 030 5109179 – Tel. 030 2400330 (5 linee con r.a.)
Sito web: www.summer.it – E-mail: summer@summer.it

2.6.1 PROGRAMMAZIONE GENETICA

Per quanto attiene le logiche elaborative dell'algoritmo di "data mining", il modello matematico-computazionale implementato dal Sistema si basa su un paradigma di programmazione genetica.

La programmazione "genetica" mutua, in ambito computazionale, tecniche di calcolo ispirate al modello dell'evoluzione naturale al fine di individuare soluzioni tendenzialmente ottimali a problemi che, per la complessità dello spazio di ricerca, non potrebbero essere risolti in tempi accettabili tramite metodi esatti.

Da un punto di vista astratto, i paradigmi di "problem solving evolutivo" rappresentano lo spazio delle soluzioni alla stessa stregua di una "popolazione di individui" per i quali risulta definita una funzione di "fitness" che ne misura la distanza dall'ottimo.

L'Algoritmo Genetico (d'ora innanzi abbreviato in AG) genera, attraverso regole di aggregazione che garantiscono la possibilità di esplorare l'intero spazio e criteri di selezione che favoriscono la sopravvivenza degli individui "migliori", taluni sottoinsiemi di tale popolazione, detti "Tribù", caratterizzati da una "fitness" che, di generazione in generazione, tende ad aumentare.

L'algoritmo viene arrestato quando, dopo un certo numero di iterazioni, è possibile ritenere che la "fitness" dell'individuo "migliore" non subirà ulteriori significativi incrementi.

Lo schema di AG da noi adottato include le seguenti fasi elaborative:

generazione casuale di una "tribù" iniziale di t individui;
ripetizione dei seguenti step fino al raggiungimento della condizione di arresto:

calcolo della "fitness" S_i per tutti gli individui che, in questo modo, possono essere ordinati come riportato in Figura 1;

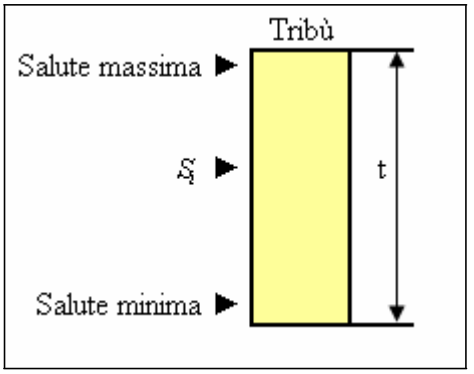

Figura 1

Reg. Soc. Trib. Brescia n. 40188 – Codice Fiscale: 07073330156 - Partita IVA: 03051110173 – C.C.I.A.A. 314779
Sede: Viale Italia, 1 – 25126 Brescia – Fax 030 5109179 – Tel. 030 2400330 (5 linee con r.a.)
Sito web: www.summer.it – E-mail: summer@summer.it

valutazione della "fitness" media $\lceil_T = \frac{\sum_{i=1}^{t} S_i}{t}$ e del relativo scarto quadratico medio $\lceil_T = \sqrt{\frac{\sum_{i=1}^{t} (S_i - \lceil_T)^2}{t}}$ della "Tribù";

suddivisione della "tribù" in due classi di individui aventi "fitness" rispettivamente superiore e non superiore (individui "vulnerabili") alla media $\lceil_T$, così come rappresentato in figura 2;

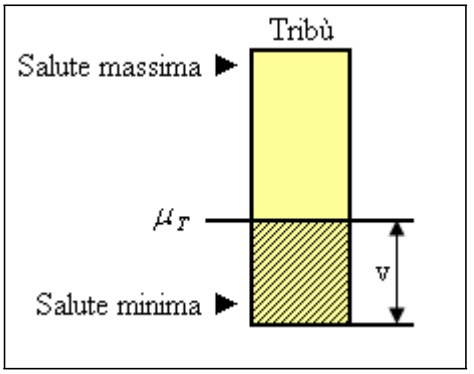

Figura 2

estrazione di $\frac{v}{2}$ coppie casuali di individui distinti, alle quali verrà applicata la seguente regola:

- se la "fitness" di entrambi gli individui non rientra nell'intervallo $[\lceil_T - \lceil_T, \lceil_T + \lceil_T]$ evidenziato in figura 3, la coppia verrà ignorata;
- diversamente, essa sarà sottoposta a "cross over", generando due nuovi individui che sostituiranno una coppia estratta casualmente dall'insieme dei soggetti "vulnerabili";

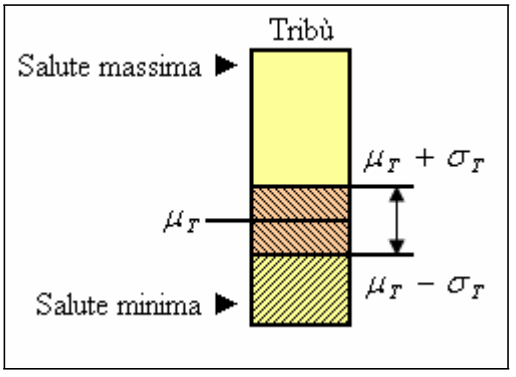

Figura 3

Reg. Soc. Trib. Brescia n. 40188 – Codice Fiscale: 07073330156 - Partita IVA: 03051110173 – C.C.I.A.A. 314779
Sede: Viale Italia, 1 – 25126 Brescia – Fax 030 5109179 – Tel. 030 2400330 (5 linee con r.a.)
Sito web: www.summer.it – E-mail: summer@summer.it

applicazione della funzione di "mutazione" agli individui "vulnerabili" che non sono stati sostituiti a seguito di "cross over".

Ogni iterazione degli step computazionali sopra descritta determina la generazione di una nuova "tribù" a partire dalla precedente.

Le regole di "mutazione" e "cross over" sono di seguito descritte:

"mutazione": trattasi di selezionare un "gene" G dal patrimonio genetico dell'individuo e di derivarne il complementare $\overline{G}$, così come rappresentato in figura 4;

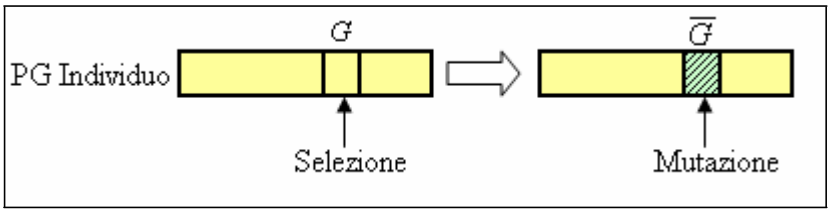

Figura 4

Il complementare di un "gene" G è ottenuto applicando una transizione "circolare" definita sul suo dominio e codificata in un'apposita tabella. Se il dominio del "gene" è binario, vale la classica definizione di complementare booleano, rappresentato in figura 5, per il quale risulta verificata la proprietà $\overline{\overline{G}} = G$.

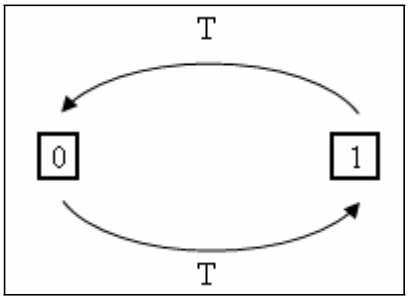

Figura 5

Reg. Soc. Trib. Brescia n. 40188 – Codice Fiscale: 07073330156 - Partita IVA: 03051110173 – C.C.I.A.A. 314779
Sede: Viale Italia, 1 – 25126 Brescia – Fax 030 5109179 – Tel. 030 2400330 (5 linee con r.a.)
Sito web: www.summer.it – E-mail: summer@summer.it

Nel caso di dominio avente cardinalità maggiore di due, ad esempio l'insieme {A,B,C,D}, la nozione di transizione binaria è naturalmente generalizzata, così come riportato in figura 6.

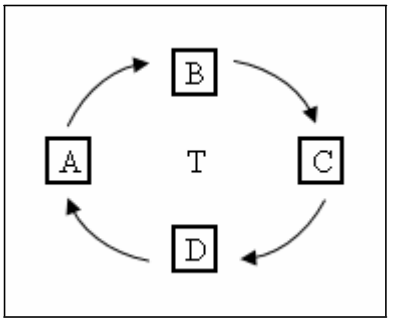

Figura 6

Determinare il complementare di un gene significa, pertanto, modificarne il valore (altrimenti detto "allele") mediante il carattere immediatamente successivo in base alla sequenza circolare indotta dalla transizione. Rileviamo che, per una transizione non binaria, non vale la proprietà $\overline{\overline{G}} = G$.

Il complementare di una sequenza di geni (utilizzato nella fase di "cross over" di seguito descritta) risulta definito, per estensione, in base alle modalità illustrate in figura 7.

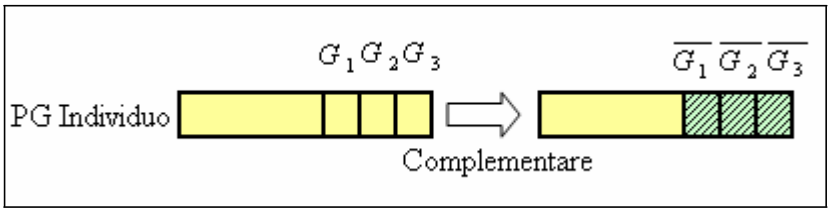

Figura 7

Reg. Soc. Trib. Brescia n. 40188 – Codice Fiscale: 07073330156 - Partita IVA: 03051110173 – C.C.I.A.A. 314779
Sede: Viale Italia, 1 – 25126 Brescia – Fax 030 5109179 – Tel. 030 2400330 (5 linee con r.a.)
Sito web: www.summer.it – E-mail: summer@summer.it

"cross over": trattasi di una funzione che, ispirandosi al meccanismo genetico del taglio e dell'incrocio del DNA, genera, a partire da due individui "genitori", una coppia di "figli" in base alla regola descritta in figura 8.

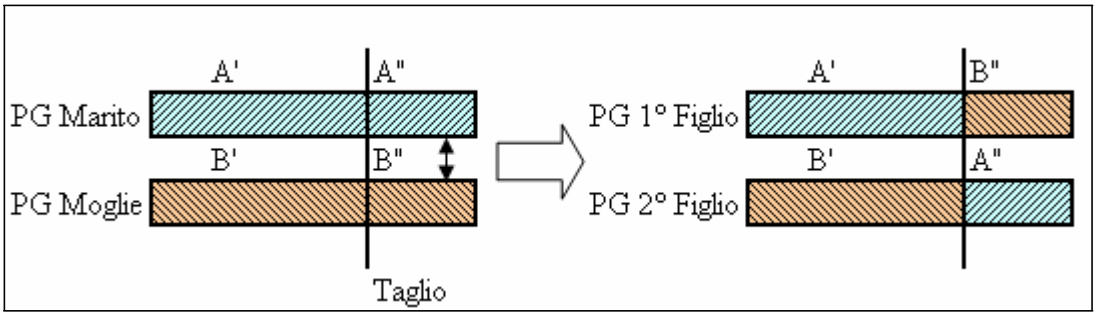

Figura 8

Nell'ambito dell'AG illustrato, ogni genitore può esibire una "fitness" S_i maggiore o, alternativamente, minore o uguale rispetto alla salute media $\lceil_T$ della "tribù".

Definiremo, pertanto, quattro regole di produzione:

- 1° *caso* $\{(S_{Marito} > \lceil_T) \quad (S_{Moglie} > \lceil_T)\}$ $\Rightarrow$ $\begin{cases} 1°\,Figlio = (A' \quad B'') \\ 2°\,Figlio = (B' \quad A'') \end{cases}$;
- 2° *caso* $\{(S_{Marito} > \lceil_T) \quad (S_{Moglie} \leq \lceil_T)\}$ $\Rightarrow$ $\begin{cases} 1°\,Figlio = (A' \quad B'') \\ 2°\,Figlio = (\overline{B'} \quad A'') \end{cases}$;
- 3° *caso* $\{(S_{Marito} \leq \lceil_T) \quad (S_{Moglie} > \lceil_T)\}$ $\Rightarrow$ $\begin{cases} 1°\,Figlio = (\overline{A'} \quad B'') \\ 2°\,Figlio = (B' \quad \overline{A''}) \uparrow \end{cases}$;
- 4° *caso* $\{(S_{Marito} \leq \lceil_T) \quad (S_{Moglie} \leq \lceil_T)\}$ $\Rightarrow$ $\begin{cases} 1°\,Figlio = (\overline{A'} * \overline{B''}) \leftrightarrow \\ 2°\,Figlio = (B' * \overline{A''}) \uparrow \leftarrow \end{cases}$.

Reg. Soc. Trib. Brescia n. 40188 – Codice Fiscale: 07073330156 - Partita IVA: 03051110173 – C.C.I.A.A. 314779
Sede: Viale Italia, 1 – 25126 Brescia – Fax 030 5109179 – Tel. 030 2400330 (5 linee con r.a.)
Sito web: www.summer.it – E-mail: summer@summer.it

2.6.2 FORMALIZZAZIONE DEL PROBLEMA

Il problema di stima di una funzione non nota a partire da un set di evidenze storiche censite presso un "data mart" è riconducibile ad un generico problema di classificazione aggredibile mediante tecniche genetiche.

Il "data mart" definisce un sottoinsieme $O \quad D_1 - .. - D_r$ di *r*-uple ordinate ed implicitamente una funzione $f : O \quad Y$ che, ad ogni *r*-upla, associa la relativa "misura"[1].

In tale contesto, siano $\nabla = \{P_1,.., P_n\}$ e $\Pi = \{Y_1,.., Y_m\}$ "partizioni" di O e di Y rispettivamente[2].

Ad ogni riga della matrice $n - m$ delle frequenze $p_{i,j} = \frac{\left| f(P_i)) \; Y_j \right|}{\left| P_i \right|}$ risulta associata una funzione "entropia" $E_i = \sum_{j=1}^{m} p_{i,j} \log \frac{1}{p_{i,j}}$. Assegnata la partizione Π, è pertanto possibile attribuire un valore di "entropia media" alla partizione ∇:

$$E_{\nabla} = \frac{\sum_{i=1}^{n} \left| P_i \right| E_i}{\left| O \right|} = \frac{\sum_{i=1}^{n} \sum_{j=1}^{m} \left| f(P_i)) \; Y_j \right| \log \frac{\left| P_i \right|}{\left| f(P_i)) \; Y_j \right|}}{\left| O \right|}.$$

Come si evince, la funzione di "entropia media" misura il grado di "disordine" della partizione *P* rispetto alla partizione *Y*, vale a dire la tendenza della funzione *f* a non assumere valori appartenenti alla stessa classe di *Y* all'interno delle varie classi di *P*. L'opposto di tale "entropia media" esprime, pertanto, il grado di prevedibilità della funzione *f* rispetto alla partizione *Y* mediante la partizione *P* utilizzata.

[1] La relazione *f* potrebbe non essere formalmente di tipo funzionale qualora ad una stessa *r*-upla risultassero associate immagini diverse. Tuttavia, possiamo assumere che il "data mart" codifichi una funzione e che le eventuali "violazioni" della regola di "univocità" delle immagini siano imputabili al "rumore" di fondo che provoca distorsioni rispetto ai valori. L'obiettivo dell'algoritmo di "data mining" è appunto quello di individuare una relazione funzionale a prescindere da tale rumore.

[2] Una famiglia $\nabla = \{P_1,.., P_k\}$ di sottoinsiemi di O è detta "partizione" se risultano verificate le seguenti asserzioni: a) $i : P_i \quad O$, b) $i \quad j : i \quad \neg j \; \Rightarrow \; P_i) P_j = \quad$ e c) $\bigcup_i P_i = O$.

Reg. Soc. Trib. Brescia n. 40188 – Codice Fiscale: 07073330156 - Partita IVA: 03051110173 – C.C.I.A.A. 314779
Sede: Viale Italia, 1 – 25126 Brescia – Fax 030 5109179 – Tel. 030 2400330 (5 linee con r.a.)
Sito web: www.summer.it – E-mail: summer@summer.it

Supponendo di disporre, per ogni "dimensione" D_h del "data mart", di l_h partizioni ∇_1^h ,.., $\nabla_{l_h}^h$, sono assegnati univocamente i "filtri" $\overline{\nabla}_1^h$,.., $\overline{\nabla}_{l_h}^h$ tali che, per qualsiasi sottoinsieme $A \subseteq O$, risulti $\overline{\nabla}_k^h(A) = \left\{\left\{o \in A : o_h = P_{k,1}^h\right\}, .., \left\{o \in A : o_h = P_{k,q_k^h}^h\right\}\right\}$.

Definiamo, in tale contesto, una "partizione ad albero" dell'insieme O in modo ricorsivo mediante le seguenti asserzioni:

$\{O\}$ è una "partizione ad albero";
se $\{P_1,.., P_k\}$ è una partizione ad albero e $\overline{\nabla}_{h1},.., \overline{\nabla}_k$ sono filtri, allora $\left\{\bigcup_{s=1}^{k} \overline{\nabla}_s^h(P_s)\right\}_f^{\leq\infty} * \{P_{k\ h+1},.., P_k\}$ è una "partizione ad albero".

Una siffatta "partizione" può essere rappresentata tramite una struttura ad albero ("albero di classificazione"), associando ad ogni nodo un "filtro" e, ad ogni ramo che si diparte dal nodo, uno dei sottoinsiemi che lo compongono.

Si noti che la medesima "partizione ad albero" può essere ottenuta attraverso sequenze di applicazione dei "filtri" differenti. Ne consegue che alberi diversi possono generare la stessa "partizione". Risulta quindi naturalmente assegnata una relazione di equivalenza tra gli alberi. Assegnata una "partizione" di *Y*, ad ogni "partizione ad albero" di *O* risulta associato un valore di "entropia media" sulla base della definizione sopra fornita. L'opposto di tale valore esprime, come già detto, il grado di prevedibilità della funzione *f*, rispetto alla partizione *Y*, mediante la "partizione ad albero" individuata.

Il problema di stimare la funzione *f* sulla base del "data mart" selezionato può, pertanto, essere ridotto all'individuazione di una "partizione ad albero" che, assegnati un insieme finito di filtri ed una "partizione" della "misura" (cioè una "interpretazione" del "data mart"), minimizzi la funzione "entropia media".

Tale problema è aggredibile mediante un approccio genetico. La funzione di "fitness" può infatti essere definita dall'opposto dell'"entropia media", mentre il DNA degli individui può essere rappresentato dalla sequenza di "filtri" all'interno degli "alberi di classificazione" che li rappresentano. Ogni nodo può essere, pertanto, paragonato ad un "gene" i cui "alleli" sono costituiti dai possibili filtri. Una "sequenza di geni" corrisponde ad un "sottoalbero di classificazione".

Sulla scorta di tali corrispondenze, il Sistema implementa uno schema evolutivo analogo a quello illustrato al punto 2.6.1 "Programmazione Genetica". Nello specifico, valgono le seguenti regole:

Reg. Soc. Trib. Brescia n. 40188 – Codice Fiscale: 07073330156 - Partita IVA: 03051110173 – C.C.I.A.A. 314779
Sede: Viale Italia, 1 – 25126 Brescia – Fax 030 5109179 – Tel. 030 2400330 (5 linee con r.a.)
Sito web: www.summer.it – E-mail: summer@summer.it

assegnate la massima profondità *p* degli "alberi di classificazione" ed il numero *v* di iterazioni dell'"esperimento", il software viene avviato *v* volte in corrispondenza di tutte le profondità comprese tra 2 e *p*. I Risultati delle $v \cdot (p-1)$ elaborazioni vengono storicizzati ai fini della consultazione;

ogni individuo della "tribù" casuale iniziale è costituito da un albero i cui nodi corrispondono ad una "partizione" di una delle "dimensioni" del "data mart" (estratta tra quelle censite nell'ambito dell'"interpretazione" di riferimento) e i cui rami sono associati ognuno ad un sottoinsieme di tale "partizione". Al fine di garantire una maggiore duttilità della struttura dati, il Sistema assume che le "partizioni" abbiano tutte la stessa cardinalità, completandole, eventualmente, con un numero congruo di sottoinsiemi vuoti;
la "mutazione" di un nodo (esemplificata in figura 9) avviene "complementando" il filtro associato, tramite la trasformazione definita dalla chiusura circolare dell'ordinamento lessicografico delle coppie "dimensione"-"partizione";

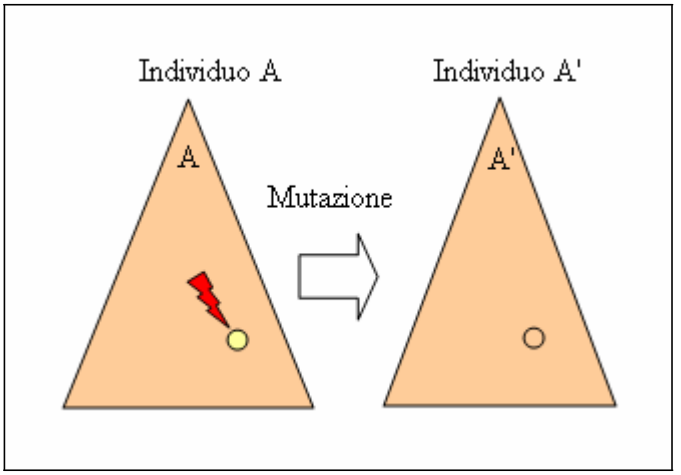

Figura 9

il "cross over" tra alberi avviene selezionando una coppia di nodi in modo casuale presso i due individui coinvolti, scambiandone i sottoalberi in essi aventi radice ed eventualmente "complementando" tali sottostrutture mediante l'applicazione già definita ai fini della "mutazione" (figura 10);

Reg. Soc. Trib. Brescia n. 40188 – Codice Fiscale: 07073330156 - Partita IVA: 03051110173 – C.C.I.A.A. 314779
Sede: Viale Italia, 1 – 25126 Brescia – Fax 030 5109179 – Tel. 030 2400330 (5 linee con r.a.)
Sito web: www.summer.it – E-mail: summer@summer.it

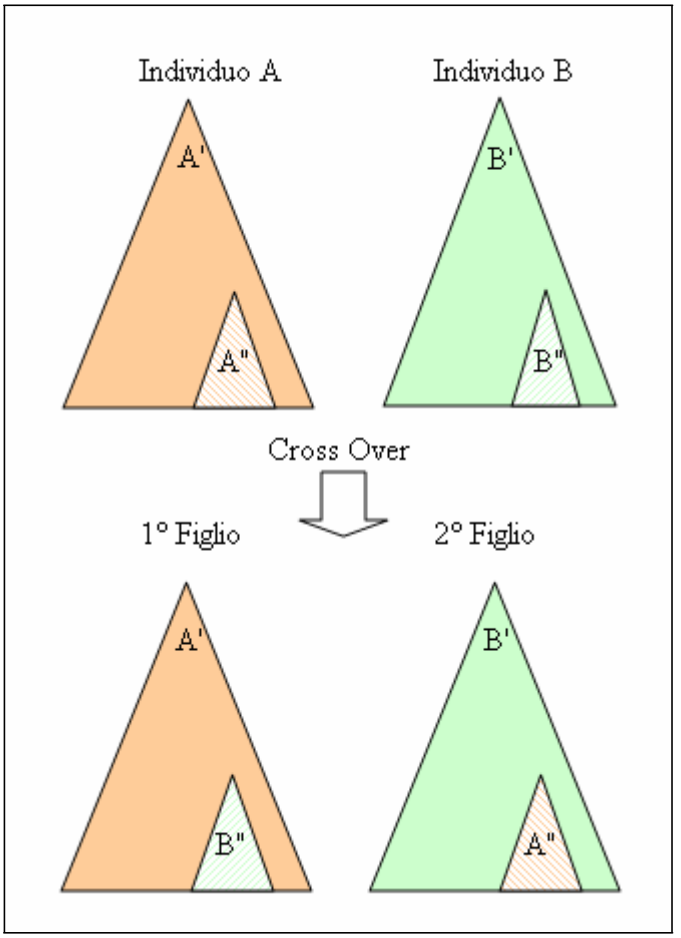

Figura 10

gli alberi generati in seguito a “mutazione” o “cross over” potrebbero presentare ridondanze dovute alla ripetizione di uno stesso filtro nell’ambito di un percorso di attraversamento. Al fine di evitare il decadimento informativo della “tribù”, il Sistema provvede, pertanto, a “normalizzare” tali tipologie di individui applicando, ai nodi ripetuti nello stesso “path”, l’operatore di mutazione;

la funzione di “fitness” di ogni albero della tribù all’inizio di un’epoca è calcolata, ai fini delle selezioni operate dall’algoritmo, presso il sottoinsieme di “training” dei “fatti” appartenenti al “data mart”. Una volta generati i nuovi individui (tramite “cross over”, “mutazione” e “normalizzazione”), le relative “fitness” vengono invece misurate nell’ambito del set dei “fatti” rimanenti, detto sottoinsieme di “test”. L’individuo migliore viene confrontato con l’ottimo derivante dalle epoche precedenti e, qualora esibisca una “fitness” maggiore, viene memorizzato in suo luogo. Tale impostazione consente all’algoritmo di evitare problemi di “overfitting” e di ridurre l’impatto di eventuali decadimenti concentrati in epoche tardive;

Reg. Soc. Trib. Brescia n. 40188 – Codice Fiscale: 07073330156 - Partita IVA: 03051110173 – C.C.I.A.A. 314779
Sede: Viale Italia, 1 – 25126 Brescia – Fax 030 5109179 – Tel. 030 2400330 (5 linee con r.a.)
Sito web: www.summer.it – E-mail: summer@summer.it

al fine di garantire che ogni nodo rappresenti una "classe" significativa, il Sistema provvede, immediatamente dopo la generazione di un nuovo individuo, ad effettuare un'attività di "razionalizzazione", selezionando le foglie dell'albero caratterizzate da una cardinalità insufficiente ad assicurarne la rappresentatività della media e "riassorbendole" nel nodo padre, congiuntamente ai sottoalberi aventi radice nei nodi fratelli;

l'esecuzione di un "run" dell'"esperimento" termina al raggiungimento della condizione di arresto, definita in base ai criteri di configurazione;

per ogni "classe" corrispondente ad un nodo dell'albero isolato, il Sistema determina il valore "sintetico" della funzione indagata in base alla seguente sequenza di calcolo:

trasformazione numerica dei sottoinsiemi della "partizione" inerente la variabile dipendente, tramite i "rappresentanti" specificati in fase di configurazione dell'"esperimento";

attribuzione, ad ognuno di tali sottoinsiemi, di una "probabilità di appartenenza" espressa come rapporto tra la relativa cardinalità ed il numero di "fatti" contenuti nel "data mart";

applicazione dell'operatore di aggregazione alla distribuzione di probabilità così ottenuta.

Si precisa che, qualora siano stati specificati criteri di amplificazione, preliminarmente all'avvio di ogni esperimento, il Sistema provvede ad esplodere i "fatti" del "data mart" in base alle regole indicate, realizzando di fatto una forma di proiezione implicita interna.

Reg. Soc. Trib. Brescia n. 40188 – Codice Fiscale: 07073330156 - Partita IVA: 03051110173 – C.C.I.A.A. 314779
Sede: Viale Italia, 1 – 25126 Brescia – Fax 030 5109179 – Tel. 030 2400330 (5 linee con r.a.)
Sito web: www.summer.it – E-mail: summer@summer.it

2.7 RAPPRESENTAZIONE E CONSULTAZIONE DEI RISULTATI

Gli $v \cdot (p-1)$ alberi "soluzione" elaborati a seguito dell'esecuzione di un esperimento vengono memorizzati sottoforma di file XML, i cui nodi risultano identificati:

dall'elemento del filtro del nodo padre che li ha generati;
dall'insieme dei nodi figli

e caratterizzati dai seguenti attributi:

popolazione della "classe" isolata dal nodo;
distribuzione di probabilità rispetto alla "misura" all'interno di tale "classe";
valore della stima della "funzione" in corrispondenza del nodo;
valori degli indicatori statistici relativi all'"entropia", all'"entropia media" del sottoalbero, alla "media", alla "mediana", alla "varianza", al "minimo" e al "massimo" che caratterizzano la distribuzione di probabilità associata alla "misura" all'interno della "classe".

Una specifica funzione consente di:

attraversare i vari percorsi radice-foglia dell'albero navigando il tracciato XML;
visualizzare l'andamento della "funzione" indagata, al variare della profondità dell'"albero di classificazione" e dei relativi nodi, tramite un istogramma che rappresenti le stime calcolate e le relative deviazioni standard;
inserire un commento circa il "risultato elaborato";
selezionare il "risultato" che s'intende utilizzare quale "soluzione" del problema.