6.14 LINEA D2 - SVILUPPO DI RETI PROBABILISTICHE DI SUPPORTO ALLE DECISIONI VITICOLTURALI ED ENOLOGICHE

Coordinatore scientifico:

Prof. Federico Stefanini - Dipartimento di Statistica – Università degli Studi di Firenze

AUTORE DEL DOCUMENTO

Prof. Federico Stefanini - Dipartimento di Statistica " G. Parenti"
Università di Firenze –
Viale Morgagni N. 59 – Firenze

1) Premessa

Il conseguimento degli obiettivi di miglioramento qualitativo agroalimentare delle produzioni dipende sia dal raggiungimento degli obiettivi specifici delle singole attività che dal livello di fruibilità sul territorio dei risultati medesimi. Il massimo impatto del progetto sul territorio si realizza quando i contenuti della ricerca trovano piena integrazione e divengono risorsa accessibile anche per la singola azienda o cantina.

A questo fine risulta utile elencare diversi livelli di fruibilità, come effettuato nell'elenco seguente:

- L1: Disseminazione in forma di report tecnico dei risultati della ricerca, eventualmente corredato da una sintesi sottoforma di *guidelines* operative.
- L2: In aggiunta a quanto previsto in L1, questo livello affianca una base dati relazionale, eventualmente interrogabile via web, contenente i dati sperimentali originali. La fruizione da parte di aziende del settore potrebbe consistere nella valutazione di risultati in condizioni sperimentali analoghe a quelle sperimentate oppure potrebbe prevedere una rianalisi statistica alla luce di esigenze peculiari dell'azienda interessata. Occorre sottolineare che una collezione di dati sperimentali in forma di strutture rettangolari separate certo contiene lo stesso ammontare complessivo di informazione ma l'accessibilità difficoltosa riduce il valore della base dati stessa, limitando di fatto la fruibilità finale.
- L3: In aggiunta a quanto elencato in L2, questo livello la fruibilità raggiunge il livello di calcolo statistico sottoforma di *script* procedurali che possono essere eseguiti in una opportuna piattaforma computazionale, ad esempio il software R, per riprodurre almeno parzialmente le analisi operate sui risultati sperimentali, ed eventualmente costituire la base di partenza da modificare per effettuare analisi similari con diverse strutture di campo. La completa fruibilità a questo livello richiede l'accesso alla base dati (L2) e un riferimento interpretativo costituito da L1.
- L4: Inevitabilmente nel livello L3 non sono contenuti quegli elementi di integrazione complessiva che, a partire dai sistemi sperimentali sottoposti a test in condizioni parzialmente controllate, riconducono i risultati sperimentali in un regime di esercizio che possiamo definire di campo, o di cantina, quindi per propria natura potenzialmente distanti dal conteso sperimentale di ricerca. Per esemplificare questo aspetto, si

consideri un fattore sperimentale definito da 3 livelli di trattamento in un conteso di prove randomizzate in condizioni controllate, cioè un'analisi standard, in cui si studiano un certo numero di trattamenti per ognuno dei tre livelli, quindi a procedere con l'analisi dei risultati. In contesto extrasperimentale potrebbe verificarsi che tale fattore sia sotto controllo, ed in tal caso sceglieremo il livello ottimale di trattamento, oppure potrebbe succedere che il fattore non sia sotto controllo, da cui la necessità di definire una distribuzione statistica che definisca la probabilità di avere uno dei tre possibili livelli di trattamento sulle unità statistiche di interesse.
In conclusione, trasferire modelli sperimentali parziali, o meglio locali, in un contesto integrato significa probabilizzare le sorgenti di incertezza coinvolte al fine di costruire un modello probabilistico integrato, ovvero una rete probabilistica, da impiegare a fini previsionali e di supporto alle decisioni in condizioni di incertezza.
A questo livello di fruibilità è richiesta la costruzione di una vera e propria API (*Application Programmino Interface*) per lo sviluppo di modelli probabilistiche integrati, uno strumento dotato di adeguata flessibilità rispetto le analisi correnti e future.

Nel seguito di questo documento saranno discusse le modalità atte al conseguimento progressivo dei livelli di fruibilità sopra discussi.

2) Metodologia statistica

L'inferenza statistica Bayesiana permette l'impiego di informazione a-priori in aggiunta all'informazione di tipo sperimentale (Bernardo e Smith, 1994, Lindey 2000). I metodi Bayesiani sono in rapida espansione anche in ambito in ambito della produzione alimentare (Boekel, et al. 2004).
Nell'inferenza Bayesiana i due tipi di informazione menzionati, a priori e campionaria, sono quantificate attraverso due distribuzioni (di densità) di probabilità, dette rispettivamente **priori** e **verosimiglianza**. I parametri del modello sono considerati a tutti gli effetti come variabili casuali.
Operativamente l'informazione contenuta nella a-priori è modificata dall'informazione presente nel campione attraverso la regola di Bayes. L'inferenza ha esito nell'ottenimento della

distribuzione del parametro condizionata alle osservazioni, ed è detta **a-posteriori** del parametro.

La regola di Bayes costituisce il motore inferenziale di tutta l'inferenza Bayesiana: sia $\{H_i\}$ una partizione finita di eventi in Ω, A un generico evento e $P[H_i]$ ed $P[A|H_i]$ le rispettive probabilità; allora segue la probabilità condizionale (od inversa):

$$P[H_i \mid A] = \frac{P[A \mid H_i]P[H_i]}{\sum_j P[A \mid H_j]P[H_j]}$$

Spesso con x si indicano i dati del campione, X la variabile casuale rispettiva definita sullo spazio campionario X, $p_\theta(x)$ è la distribuzione di X che appartiene alla famiglia $\mathsf{P}=\{p_\theta(x); \theta \in \Omega\}$ con parametro uni o multidimensionale θ appartenente a Ω spazio dei parametri, $\pi(\theta)$ è la a-priori del parametro. Quando x sia fissato, $p_\theta(x)$ rappresenta la funzione di verosimiglianza, una funzione che quantifica come cambia la probabilità (densità) del realizzarsi di x al variare di θ.

La a-priori rappresenta l'informazione extra campionaria quali informazioni tecniche, opinioni, feed-back del processo esistente, ed anche la presenza eventuale di un super-esperimento di riferimento.

La distribuzione a-posteriori è ottenuta combinando la a-priori con la funzione di verosimiglianza (regola di Bayes):

$$\pi(\theta \mid x) = \frac{\pi(\theta) \cdot p_\theta(x)}{\int_\Omega \pi(\theta) \cdot p_\theta(x) \cdot d\theta}$$

La a-posteriori fornisce una sintesi complessiva dell'informazione esistente circa il parametro.

A partire dalla a-posteriori del parametro è possibile effettuare sintesi diverse dell'informazione complessiva, ad esempio mediante la stima puntuale Bayesiana in cui un valore puntuale della distribuzione a posteriori del parametro è scelto quale valore stimato dell'incognita θ. E' pratica frequente scegliere la moda della posteriori $\pi(\theta|x)$ come stima puntuale del parametro incognito oppure la media della distribuzione a-posteriori.

In molti casi una sintesi intervallare circa $\pi(\theta|x)$ può essere ottenuta come regione $S_\alpha(x)$ di Ω che gode della proprietà probabilistica:

$$P[\theta \in S_\alpha(x) \mid x] = \int_{S_\alpha(x)} \pi(\theta \mid x) \cdot d\theta = 1 - \alpha$$

con (1-α) il **livello di confidenza Bayesiano** della regione di confidenza Bayesiana.

La preminenza del calcolo delle probabilità nell'inferenza Bayesiana porta a test statistici di forma semplice: fissata l'ipotesi nulla H: $\theta \in w \subset \Omega$ in alternativa all'ipotesi H': $\theta \in \Omega - w$, si richiede di valutare la probabilità

$$P[H \mid x] = \int_H \pi(\theta \mid x) \cdot d\theta$$

e si rifiuta l'ipotesi nulla se tale valore di probabilità risulta piccolo, ovvero più piccolo del fissato livello critico α.

Il problema di previsione statistica in ambito Bayesiano ha una naturale formulazione in termini di distribuzione predittiva (Barnett, 1973). Indicando come y il vettore di dati futuri, la previsione si realizza attraverso la distribuzione condizionata P[y | x]:

$$P[y \mid x] = \int_\Omega p_\theta(y)\pi(\theta \mid x) \cdot d\theta = \frac{\int_\Omega p_\theta(y)\pi(\theta) \cdot p_\theta(x) \cdot d\theta}{\int_\Omega \pi(\theta) \cdot p_\theta(x) \cdot d\theta}$$

la quale contiene tutta l'informazione disponibile circa y. La distribuzione sopra riportata ammette sintesi analoghe a quanto discusso per la a-posteriori.

Infine, molti problemi decisionali ammettono una trattazione statistica quantitativa formale attraverso la definizione di funzioni di utilità per valori futuri di variabili o per certi parametri del sistema considerato. In tali casi la massimizzazione di quantità collegate a tali funzioni permette di definire decisioni ottime rispetto i criteri stabiliti.

Nell'inferenza Bayesiana il calcolo delle distribuzioni condizionate e marginali può presentare serie difficoltà operative. Inoltre, per sostanziare il contenuto informativo a-priori è necessario un dialogo non superficiale tra esperto di statistica ed esperto nel dominio del problema. L'impiego dell'inferenza Bayesiana in sistemi stocastici complessi altamente strutturati ha stimolato lo sviluppo delle reti probabilistiche (Jensen 1996, Cowell et al., 1999, Neapolitan, 2003).

La formulazione di reti probabilistiche sposta l'attenzione sulle variabili che interessano lo specialista nel dominio del problema applicativo, liberandolo almeno in parte dalla necessità di esprimere l'informazione a priori direttamente in forma quantitativa. Se alcuni requisiti tecnici sono soddisfatti, in aggiunta a strumenti grafici di rappresentazione della conoscenza è

possibile ottenere un motore computazionale che in maniera estremamente efficiente procede alle operazioni di condizionamento e marginalizzazione, le operazioni base dell'inferenza Bayesiana.

Il recente sviluppo di *shell* grafiche per lo sviluppo di reti Bayesiane ha portato a tecnologia delle reti Bayesiane alla portata di specialisti non statistici appartenenti a campi che spaziano dalla Biologia molecolare sino all'ambito agronomico (Neapolitan, 2003)

3) Obiettivi del programma di ricerca ovvero " Reti probabilistiche di supporto alle decisioni operative"

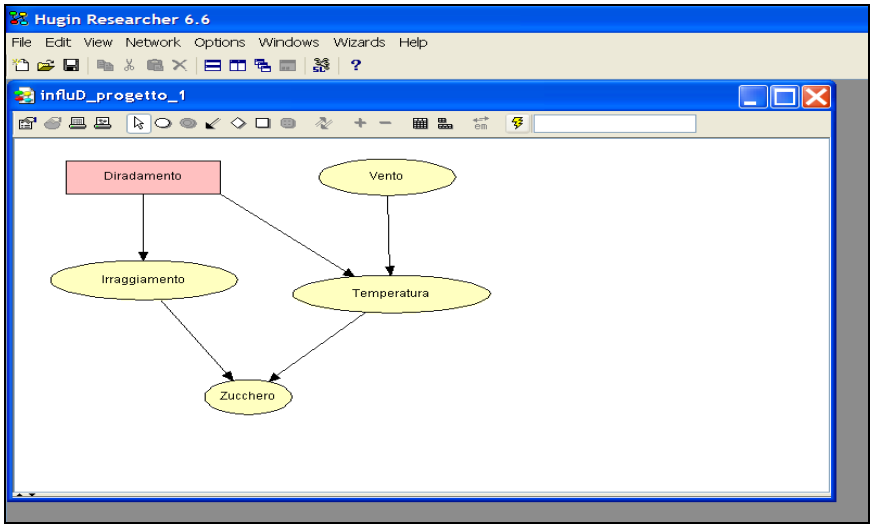
La rappresentazione di un sistema stocastico strutturato può essere effettuata ricorrendo alle reti probabilistiche. Qualora l'interesse sia diretto a fornire uno strumento di supporto alle decisioni, è possibile introdurre nella rete nodi decisionali e studiare l'effetto delle possibili decisioni sulle variabili di interesse.

Rimandiamo il lettore interessato ai dettagli statistici al volume di Neapolitan (2004, Learning Bayesian Networks, Prentice Hall), ed introduciamo un caso di studio particolarmente semplice se confrontato con il progetto intero. Tuttavia esso rappresenta uno strumento non banale per illustrare le potenzialità operative dell'approccio.

Si considerino le seguenti variabili sperimentali: Diradamento (S/N), Vento (-1,0,1, dopo stanrdardizzazione), Temperatura (-1,0,1, dopo standardizzazione), Irraggiamento (-1,0,1, dopo standardizzazione), Zucchero (0,1,2,3, dopo standardizzazione).

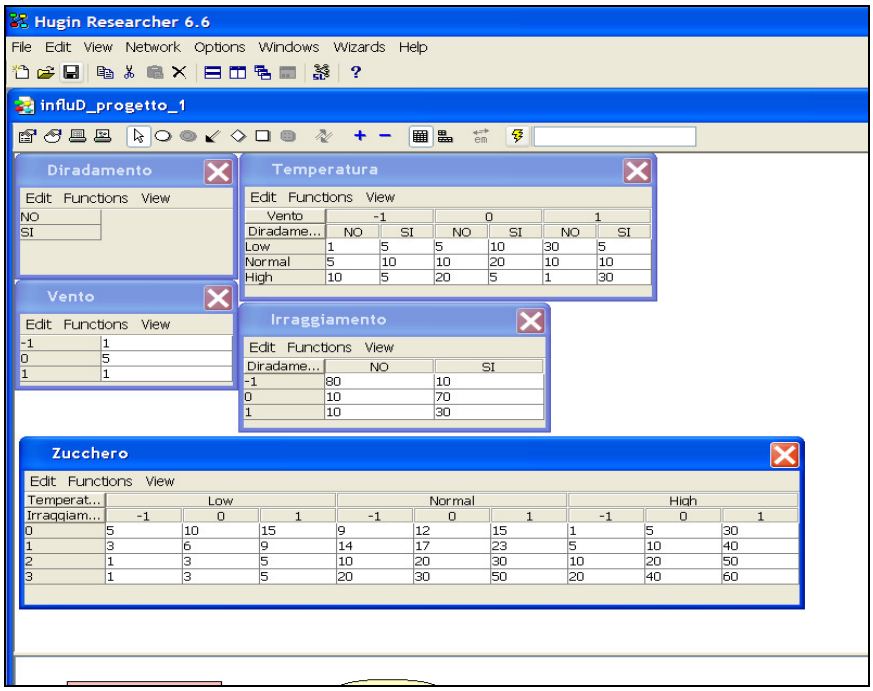
In Figura 1 è riportata una rete probabilistica che ipotizza un certo insieme di relazioni tra le variabili menzionate. Le frecce indicano la presenza di una relazione causale oppure, più debolmente, indicano la rilevanza delle variabili genitore (originatrici delle frecce) per la definizione della distribuzione di probabilità della variabile figlio (destinazione delle frecce di tali genitori). Le relazioni sono ottenute nella fase di auditing di progetto.

FIGURA 1



Dopo avere formulato la rete di dipendenze (rilevanze) tra variabili, è necessario completare la specificazione indicando per ogni possibile valore assunto dalle variabili genitore quale è la distribuzione di probabilità condizionata per la variabile figlio (FIGURA 2). Le tabelle di probabilità condizionate sono ottenute nella fase di auditing di progetto.

FIGURA 2

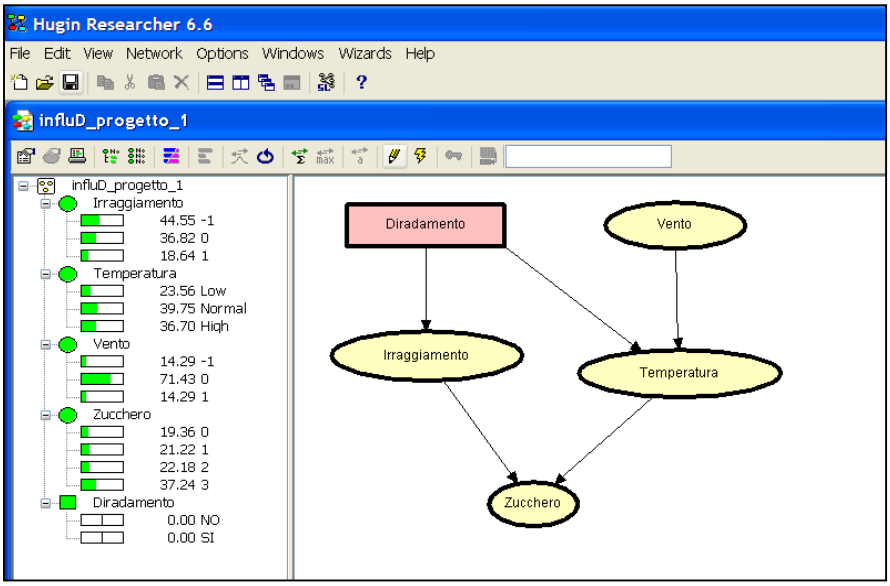


Ultimata la fase di elicitazione (auditing d progetto) si può procedere in linea teorica all'utilizzo della rete probabilistica, a condizione che non vi siano relazioni tra variabili o valori

di probabilità incerti. Se l'incertezza rimasta dopo l'auditing è sostanziale, diviene obbligatorio integrare l'informazione a priori disponibile con i risultati ottenuti con esperimenti di campo. Tecnicamente si indica questa fase dello studio come parameter learning e structural learning,

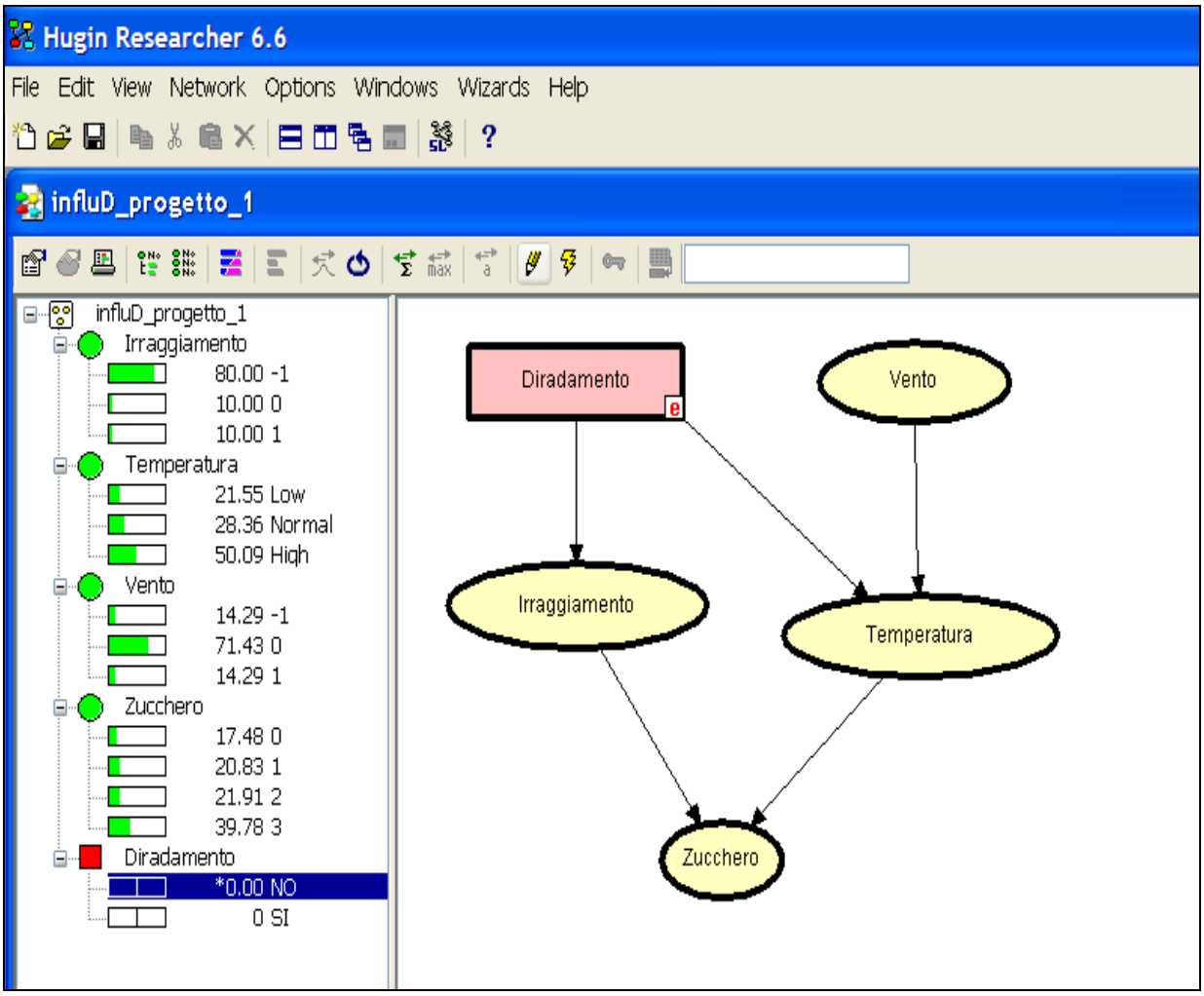
Assumiamo nel seguito che non vi siano incertezze sostanziali. In FIGURA 3 è riportata la distribuzione marginale delle singole variabili della rete prima di inserire le evidenze osservate e prima di prendere le decisioni. Le barre verdi a sinistra indicano il valore di probabilità marginale.

FIGURA 3



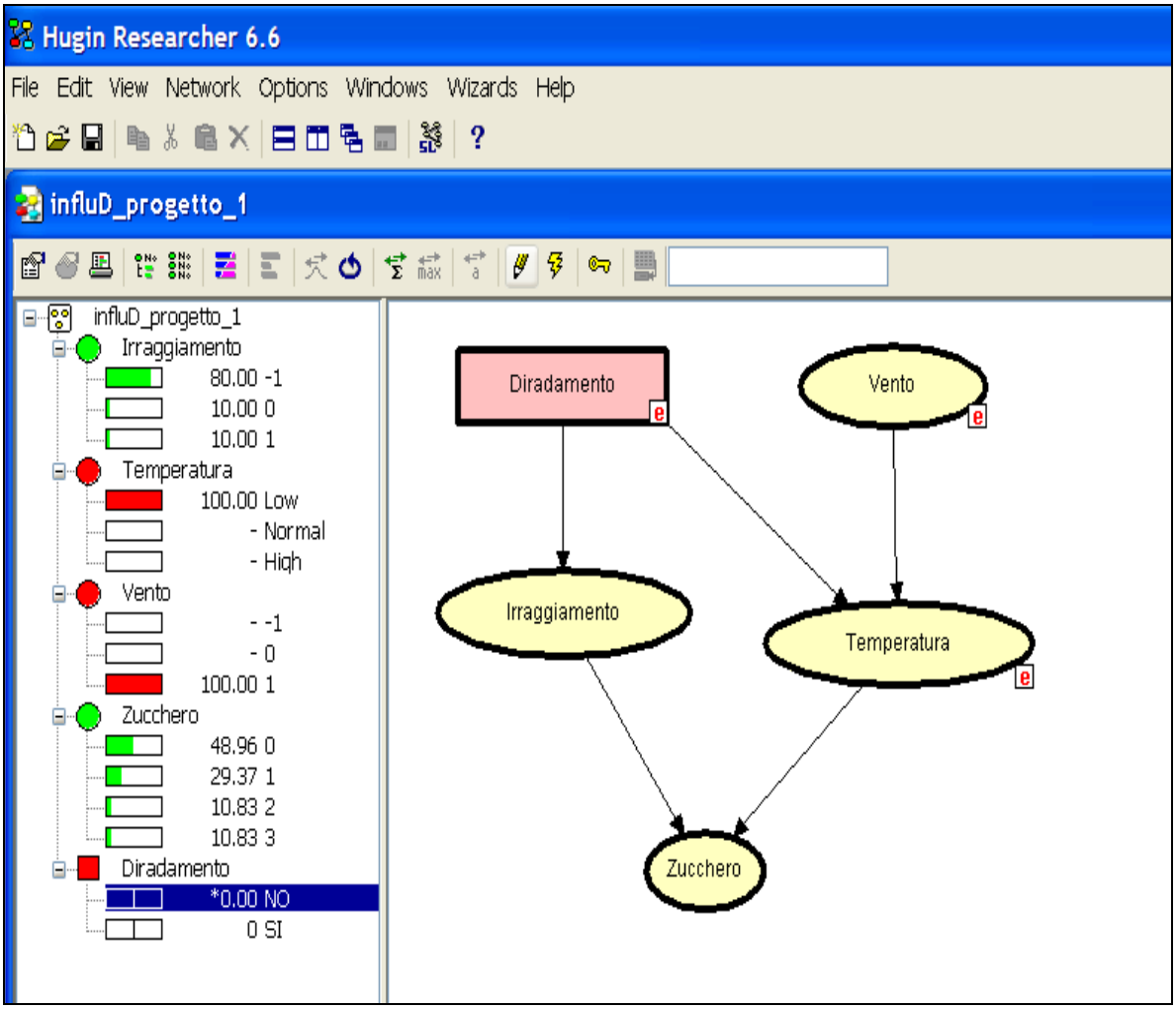
Impostando il valore decisionale NO per la variabile Diradamento (FIGURA 4), la rete propaga la decisione effettuata e modifica i potenziali probabilistici trasferendo gl effetti della decisione in tutta la rete. La distribuzione marginale delle variabili rimanenti, condizionatamente alla decisione operata cambia.

FIGURA 4



In ambito applicativo vi possono essere ulteriori evidenze sperimentali che possono essere introdotte nella rete probabilistica. In FIGURA 5, in aggiunta alla decisione sul Diradamento sono state inseriti valori di vento e temperatura (da cui il colore rosso e la letterina e).
Anche in questo caso le distribuzioni marginali delle variabili rimanenti condizionatamente alla decisione operata ed ai valori osservati di certe variabili (le evidenze) è cambiata

FIGURA 5



Discussione

Le reti probabilistiche si possono considerare concettualmente uno sviluppo dei diagrammi causa effetto noti entro discipline quali il controllo statistico dei processi ed i metodi statistici per la gestione della qualità. L'estensione non è triviale in quanto anche relazioni non causali sono ammesse e perché tali relazioni sono formulate anche in maniera quantitativa. Un altro

parente metodologico è sicuramente la tecnica della *path analysis* di S. Wrigth, rispetto alla quale la principale differenza è rappresentata dall'introduzione di variabili decisionali e dalla sola presenza di variabili casuali discrete.

Una rete probabilistica DAG con variabili gaussiane cattura le informazioni disponibili nella *path analysis*. Una naturale estensione della *path analysis* di S. Wrigth è rappresentata dall'introduzione di alcune variabili come discrete, pur escludendo nodi decisionali. La classe dei modelli così definiti è detta *mixed probabilistic networks*, la quale modella indipendenze condizionate per distribuzioni Conditionally Gaussian.

Le reti probabilistiche supportano il ragionamento causale qualora gli archi dei grafi indichino informazione sull'esistenza di un legame causale. L'inferenza statistica in tal caso non basta all'identificazione di tali relazioni, almeno senza ricorre ad esperimenti randomizzati in condizioni controllate. Tuttavia, strumenti statistici di inferenza e *machine learning* consentono di identificare spazi di ipotesi attraverso i grafi PDAG, che possono suggerire ipotesi da saggiare sperimentalmente.

L'impiego operativo degli strumenti illustrati in questa sezione richiede una calibrazione assolutamente non banale. Intuitivamente risulta chiaro che esiste sempre la possibilità di unire con frecce le variabili di oggetto di interesse ma la reale portata applicativa della rete probabilistica formulata dipende totalmente dalla sua capacità di ritrarre le caratteristiche salienti del sistema oggetto di studio.

Per questa motivazione è indispensabile partire con un'accurata azione di auditing di progetto, non solo per ottenere tutta l'informazione a priori disponibile, ma anche per calibrare la risultante rete in modo che vi sia il minor numero possibile di incongruenze. L'operazione presenta un certo grado di complessità già a partire da qualche decina di variabili.

Vi sono poi un certo numero di ulteriori difficoltà metodologiche. Tutte le variabili della rete probabilistica menzionata sono discrete, ovvero quelle originariamente continue devono essere rese discrete. Sebbene questo passo operativo potenzialmente comporti perdita di informazione o, più comunemente, influenzi la qualità delle inferenze effettuate, d'altra parte consente di effettuare i calcoli in maniera altamente efficiente.

In secondo luogo la contestualizzazione è critica. Alcune variabili si riferiscono ad intervalli temporali altre allo stato del sistema in un certo istante temporale. La scala di ogni variabile richiesta per la formulazione di una rete probabilistica realmente utile, o più in generale la

trasformazione delle variabili originali nello spazio del problema, possono non essere banali. Si considerino a tale riguardo misure in tempo continuo operate da sensoristica di campo.
Infine, pur considerando che le reti probabilistiche siano formulate proprio per riconoscere la natura modulare di sistemi complessi altamente strutturati, rimane da trattare appropriatamente l'analisi di dati sperimentali di campo, operazione non banale quando i livelli di osservazione si spostano dal micro (livello molecolare) al macro (ad esempio parcelle di campo).

4) Tecnologie e risorse computazionali

Le principali tecnologie adottate in questa proposta sono accessibili a costo zero sia da parte degli statistici coinvolti che da parte dell'utente finale.
A livello L1, in aggiunta a formati nativi (a volte proprietari) di preparazione dei documenti si prevede di uniformare in formato elettronico HTML e PDF i documenti prodotti dal gruppo di elaborazione statistica. Parte dell'elaborazione statistica sarà svolta in linguaggio statistico S ricorrendo al software multi piattaforma R, progetto *open source* il cui codice può essere liberamente scaricato via internet (http://www.r-project.org).
A livello L2, la tecnologia di riferimento per la costruzione di una base dati integrata potrebbe essere costituita dal database relazione MySQL (http://www.mysql.com), da installare come server su una personal computer di fascia media-alta a questo scopo dedicato. Fisicamente il computer potrebbe essere inserito nella *intranet* del Dipartimento di Statistica "G.Parenti", Firenze, in modo da beneficiare della presenza di protezioni quali un *firewall* professionale ed un antivirus a livello di *firewall*. E' importante sottolineare che la tecnologia R-based e MySQL sono in grado di dialogare perfettamente.
La realizzazione del livello L3 richiede un linguaggio di *scripting* computazionalmente completo. I moderni software di statistica, quale R, STATA, Splus, SAS (ed altri) contengono *facilities* di automatizzazione ed anche veri e propri linguaggi di programmazione per l'analisi statistica dei dati. Ad oggi, il progetto R sembra il solo ad unire la filosofia *open source* e ad avere contemporaneamente realizzato un vera risorsa statistica estesa, ovvero di avere superato il livello di mera prova di concetto per la fattibilità. Pertanto, si propone l'impiego estensivo del linguaggio statistico S in ambiente R per l'elaborazione dei dati. E' di estrema importanza

sottolineare che la stesura degli script di elaborazione può essere convenientemente abbinata ad output in HTML, ma che richiede la massima cura a livello di commenti e documentazione del codice perché possa effettivamente costituire un output del progetto.

Infine, per il livello di fruibilità L4 richiede un vero e proprio linguaggio di programmazione per sviluppare librerie di statistica che, in aggiunta alla capacità di elaborazione statistica numerica, possano potenzialmente dialogare in maniera efficiente con la base dati anche attraverso la rete internet. Per questi motivi è stato scelto il linguaggio JAVA (http://java.sun.com), che è anche una tecnologia altamente integrata e gratuitamente fruibile. In parte essa è anche accessibile in forma *open source*. Sono disponibili potenti strumenti di sviluppo che includono la piattaforma NetBeans e la piattaforma sviluppata entro il progetto *open source* Eclipse (http://www.eclipse.org). Dato che la quantità di API (librerie) Java è in crescente aumento anche per quanto riguarda l'elaborazione statistica, nella prima parte del progetto saranno valutate le risorse computazionali Java dedicate all'elaborazione statistica ed alle reti probabilistiche allo scopo di sfruttare al massimo grado le librerie esistenti.

In conclusione, le tecnologie di questa proposta di progetto sono altamente integrate e sono classificabili in larga parte come *open source*, o comunque sono accessibili attualmente gratuitamente, con un chiaro beneficio sulla sicurezza e sui costi di implementazione.

5) Cronoprogramma delle Attività

Anno - Attore	Attività	mesi	Output
1 - Stefanini	*1.1 Auditing* di progetto ed esame della documentazione in letteratura	5	Note di progetto
1 - Stefanini	1.2 Valutazione API esistenti per sviluppo modellistico integrato	1	Valutazione scritta
1- Stefanini	1.3 Definizione della base dati relazionale	1	Prototipo MySQL
1 - Stefanini	1.4 Definizione delle funzionalità di una API Java di supporto al progetto	2	Prototipi di rete probabilistica
2 - Stefanini	2.1 Coordinamento analisi statistiche	2	-
2 – n.1 contratto	2.2 Costituzione di base dati a partire da risultati di esperimento rettangolari		Base dati elettronica
2 n.1 contratto	2.3 Sviluppo e documentazione scripting in R		Codice R
2 – Stefanini e n.1 contratto	2.4 Estensione core API Java	6	Java API
3 - Stefanini	3.1 Coordinamento analisi statistiche	3	-
3 - n.1 contratto	3.2 Completamento base dati relazionale		Data base
3 – Stefanini e n.1 contratto	3.3 Sviluppo modello statistico integrato	2	Modelli statistici
3 - n.1 contratto	3.4 Implementazione del modello integrato		Codice R
3 - Stefanini e n.1 contratto	3.5 Sviluppo modelli statistici sperimentali di cantina	3	Modelli statistici
4 - Stefanini	4.1 Coordinamento analisi statistiche	3	
4 - Stefanini e n.1 contratto	4.2 Sviluppo modelli statistici sperimentali di cantina	2	-
4 - Stefanini e n.1 contratto	4.3 Modelli statistici per il panel testing enologico	2	Modelli statistici
4 - Stefanini e n.1 contratto	4.4 Consolidamento e test della rete probabilistica previsionale	3	Reti probabilistiche

Dettaglio attività del primo anno.

1.1) Incontro settimanale con leaders delle attività di progetto per approfondire la problematica di ricerca e sviluppo e dettagliare le esigenze metodologiche e statistiche. Ogni leader è tenuto ad inviare per tempo una scheda tipo come da allegato 1 (*in versione bozza*) in modo che il dialogo sia massimamente produttivo. Ad ogni incontro corrisponde una scheda informativa che integra quanto presentato come scheda.

1.2) In preparazione al lavoro di integrazione modellistica via reti Bayesiane, in questa attività si prevede di valutare sistematicamente le risorse API ed IDE richieste per la formulazione di reti probabilistiche di supporto alle decisioni agroalimentari.

1.3) In base all'elicitazione di informazioni accessorie, ai documenti e all'azione di auditing, saranno valutati scenari d'uso e formulata la struttura relazionale della base dati.

1.4) In base all'elicitazione di informazioni accessorie, ai documenti e all'azione di auditing, saranno valutate possibili implementazioni di reti probabilistiche di supporto alle decisioni, anche ricorrendo a valutazioni numeriche su prototipo.

6) Bibliografia

Barnett V., 1973, Comparative Statistical Inference, Wiley.

Bernardo J.M., Smith A.F.M., 1994, Bayesian Theory, Wiley.

Boekel, M.A.J.S. van; Stein, A.; Bruggen, A.H.C. van (Eds.), 2004, Bayesian Statistics and Quality Modelling in the Agro-Food Production Chain. Proceedings of the Frontis workshop on Bayesian Statistics and quality modelling in the agro-food production chain, held in Wageningen, The Netherlands, 1-14 May 2003, Series: Wageningen UR Frontis Series, Vol. 3

Cowell, R.G., Dawid, A.P., Lauritzen, S.L., Spiegelhalter, D.J., 1999, Probabilistic Networks and Expert Systems. Springer Verlag

Jensen, F.V., 1996, An Introduction to Bayesian Networks, Springer-Verlag, 1996.

Lindley D., 2000, The philosophy of Statistics. Journal of the Royal Statistical Society (A).

Neapolitan, R., 2003, Learning Bayesian Networks.Academic Press.

Pearl, J., 2000, Causality. Cambridge University Press

Per la tecnologia Java:
Tutorials all'indirizzo http://java.sun.com/docs/books/tutorial/
Omega Project for computing (R) all'indirizzo http://www.omegahat.org/

7) Curriculum vitae

Prof. Federico Mattia Stefanini

Il prof. Stefanini dirige un gruppo di biostatistici Bayesiani che studiano sistemi complessi, come quelli dei sistemi genetici molecolari investigati con la tecnologia *microarray* e con altri saggi molecolari. L'impiego integrato di sorgenti di informazioni organizzate in strutture gerarchiche è tra i principali scopi di questo gruppo. Reti probabilistiche Bayesiane e modelli grafici, uniti a metodi computazionali Monte Carlo, sono tra le tecniche statistiche correntemente in uso. Lo sviluppo di modelli statistici spesso include l'implementazione originale di software.

Il prof. Stefanini insegna "Statistica" presso la Facoltà di Agraria a tutti i livelli previsti. E' supervisore al programma di dottorato di ricerca in statistica applicata presso il Dipartimento di Statistica.

Il prof. Stefanini ha collaborato negli ultimi anni con Istituzioni e Centri di ricerca in Iatlia e All'Estero: Stanford Department of Biological Sciences, Morrison Institute for Population Studies, Santa Fè Institute, New Mexico. In Italia ha recentemente ultimato un progetto di agrometerologia con ARSSA regione Abruzzi; in precedenza ha collaborato con il Ministero degli Affari Esteri, con il CNR, ed attualmente con il progetto GeneExpress in Firenze.

La sua ricerca è stata finanziata tra il 2002 ed il 2005 da progetti PRIN nazionali per sviluppare tecnologie legate alle reti probabilistiche anche in ambito di modelli causali.

Alcune pubblicazioni pertinenti:

Panconesi A., Casini N, Santini A., Stefanini F.M.,1994, The influence of artificial *Seiridium cardinale* infection on the growth parameters of cypress clones (*Cupressus sempervirens*), Canadian Journal of Forest Research **25**:109-113.

Camussi A., Sari-Gorla M., Villa M., Greco R., Stefanini F.M., 1994, Effects of microclimatic variations on the estimate of relationships between RFLP markers and low hereditability traits in maize, Maydica **39**:129-132.

Cicogna M., Stefanini F.M., Camussi A. , 1994, Multivariate statistical analysis of relationships among morphological and functional traits in young beef bulls of piedmontese breed, (in italian), Zoot. Nutr. Anim. **21**:239-248.

Montanelli C., Chiari A., Chiari T., Stefanini F.M., Nascari G., 1995, Evaluation of resistance to P*seudomonas solanacearum* in potato under controlled conditions, Euphytica **81**:35-43.

Montanelli C., Stefanini F.M., Nascari G., Chiari T., Chiari A., 1995, Variability in the response to *Pseudomonas solanacearum* of transgenic lines of potato carrying a cecropin gene analogue, Potatoes Research **38**:371-378.

Stefanini F.M., Camussi A., 1997, Information in molecular profile components evaluated by a Genetic Classifier System: a case study in Picea abies Karst., Genetical Research, **70**:205-213.

Stefanini F.M., 1998, Identification of highly informative molecular profiles components using Genetic Algorithms, Santa Fe Institute, New Mexico, USA, working paper 98-05-042. http://www.ds.unifi.it/~stefanin/wpapsfi.zip

Stefanini F.M., Feldman M.W., 1999, Microsatellite loci and the origin of modern humans: a Bayesian analysis, Solomon P. Wasser (ed.), Evolutionary Theory and Processes: Modern Perspectives, Kluwer Academic Publishers, 1999.

Stefanini F.M., Camussi A., 1999, The choice of molecular profile components based on a quantitative evaluation of convenience, Biometric Letters, **36**:47-60.

Stefanini F.M., Surico G., Marchi G., 2000, Longitudinal Analysis of symptom expression in esca disease grapevines, Phytopathologia Mediterranea, **39**: 225-231.

Stefanini F.M., Feldman M.W., 2000, Bayesian Estimation of Range for Microsatellite Loci, Genetical Research, **75**:167-177.

Stefanini F.M., Camussi A., 2000, The reduction of large molecular profile to informative components using a Genetic Algorithm, Bioinformatics, **16**:923-931. http://www.ds.unifi.it/~stefanin/bioinformatics.htm

F.M. Stefanini, Microsatellite loci with range constraints: estimates of range and of divergence time between two populations. Center for Computational Genetics and Biological Modeling, Stanford University, Working Paper Series n°18, 2001.

F.M. Stefanini, Momenti di popolazione e inferenza sull'evoluzione in loci genetici microsatellite. Dipartimento di Statistica "G.Parenti", Firenze, Working Paper n° 91, 2001. http://www.ds.unifi.it/~stefanin/PCM/PCM.htm

Stefanini, F.M. (2002). A model-based normalization for cDNA microarray experiments. Atti della XLII Riunione Scientifica SIS, Milano

Corradi, F., Lago G., F.M.Stefanini,2003, The evaluation of DNA evidence in pedigrees requiring population inference, Journal of the Royal Statistical Society ser A, 166, Part 3, pp. 425–440.

Cidronali*, V. Nair, G. Collodi, J. Lewis, M. Camprini*, G. Manes*, H. Goronkin, S. Selleri*, F. Stefanini, A. Giusti, 2003, MMIC Applications of Heterostructure Interband Tunnel Devices, IEEE Transactions on Microwave Theory and Techniques, special issue.

Stefanini, F.M. 2003, The normalization of microarray data: a Bayesian graphical model. Invited paper, Congress Acta, International Biometric Society, Regione Italiana.

Camussi A. Stefanini F.M., 2005, La classificazione di cloni di pioppo con metodi Monte Carlo: le foreste casuali. Forest@ 2(2): 217-224.

Mascherini M., Stefanini F.M., 2005, M-GA: a genetic algorithm to search for the best Conditional gaussian networks, accepted for oral presentation, International Conference on Computational Intelligence for Modelling, Control and Automation, Nov. 2005

Submitted

Stefanini, F.M., 2004, Background fluorescence in cDNA microarray experiments, submitted.

C. Mavilia, M. Mascherini, V. Martineti, A. Tanini, M. L. Brandi, F.M. Stefanini, 2004, The normalization of DNA microarrays using spike controls: an additive model, submitted.