UN SISTEMA SPERIMENTALE DI ACQUISIZIONE, MEMORIZZAZIONE ED ELABORAZIONE DI INFORMAZIONI SANITARIE IN CAMPO ONCOLOGICO*

Vitoantonio Bevilacqua[1,2], Vito Santarcangelo[1], Alberto Magarelli[1] Giuseppe Oddo[1], Rocco Scaramuzzi[1], Dario Turco[3], Guido Riccio[3], Primiano Di Nauta[4]**

[1] Dipartimento di Elettrotecnica ed Elettronica – Politecnico di Bari , via Orabona, 4 70125 Bari, Italy
[2] eBIS s.r.l. Spin-Off del Politecnico di Bari – Via Pavoncelli, 139 , 70125 Bari , Italy
[3] Agilex s.r.l. Via E. Estrafallaces, 2/4, 73100 Lecce , Italy
[4] Università di Foggia – Facoltà di Economia, Via Caggese, 1 - 71100 - Foggia - Italy

**autore di riferimento - email: bevilacqua@poliba.it*

Riassunto

Nel presente lavoro si dimostra, inizialmente, la potenzialità di alcuni strumenti di apprendimento basati su algoritmi attualmente disponibili nella piattaforma WEKA di Pentaho, sempre più utilizzata sia in ambito sanitario scientifico sia in ambito professionale. Successivamente, si presenta un nuovo ambiente progettato e implementato presso il Dipartimento di Elettrotecnica ed Elettronica del Politecnico di Bari, per far fronte alla mancata disponibilità di dati multi settoriali quali quelli necessari per predire e classificare il rischio di insorgenza di malattie tumorali multifattoriali, ovvero caratterizzate da fattori di rischio su base genetica, ambientale e alimentare. Lo scopo finale del lavoro è quindi quello di affermare che, grazie alla potenzialità di strumenti inferenziali particolarmente versatili e robusti, e al loro utilizzo su una quantità significativamente rilevante, dal punto di vista statistico, di dati acquisiti anche sotto il suggerimento di esperti medici del settore, è possibile ottenere informazioni sanitarie di notevole interesse, non soltanto dal punto di vista diagnostico o prognostico, ma anche dal punto di vista gestionale ed economico.

* Sebbene il contributo sia stato riletto e condiviso da tutti gli autori, l'intero lavoro di ricerca e redazione è stato prevalentemente condotto dagli autori Vitoantonio Bevilacqua, Vito Santarcangelo e Alberto Magarelli, i co-autori Rocco Scaramuzzi e Giuseppe Oddo hanno rispettivamente contribuito alla redazione di alcune parti dei paragrafi 2 e 3. Dario Turco, Guido Riccio e Primiano Di Nauta hanno consentito il reperimento delle informazioni necessarie a comprendere la necessità di progettare e implementare l'intero sistema.

1. INTRODUZIONE

Il carcinoma del colon retto ha un'elevata incidenza nei paesi occidentali, vicina a quella del carcinoma polmonare (nel sesso maschile) e del carcinoma mammario in quello femminile. Per il tumore al colon retto così come per quello mammario è ampiamente dimostrata l'importanza di fattori costituzionali e genetici (alta frequenza in alcune popolazioni rispetto alle altre ed elevata incidenza di casi nelle stesse famiglie), nonché un'alta possibilità di successo di soluzioni chirurgiche se la diagnosi è realizzata in fase precoce. Per tutti questi motivi, le precedenti patologie hanno suscitato, negli ultimi anni notevole interesse dal punto di vista non soltanto in ambienti scientifici medici, ma anche ingegneristici e informatici. Si veda a tal proposito la produzione scientifica di alcuni degli autori di questo lavoro [1,2], alimentata anche dalla sempre maggiore attenzione di gruppi di ricerca applicata alle campagne di sensibilizzazione della popolazione, con l'obiettivo di fornire utili strumenti di acquisizione e analisi a supporto della prevenzione di base e familiare ormai necessarie per quanto riguarda il tumore al colon retto ed alla mammella. Ebbene questo articolo prende proprio le mosse da queste istanze e propone, attraverso un percorso motivato nell'articolazione dei suoi paragrafi, un sistema di elaborazione dei dati pertinenti queste tre tipologie di tumore, al fine di dimostrare l'efficacia di alcune metodologie inferenziali note come apprendimento supervisionato, quali utili sistemi in grado di fornire semplici regole ampiamente validate e riconosciute sulla base della statistica medica. Inizialmente si presentano quindi le potenzialità di WEKA, un ambiente software di Data Mining interamente scritto in Java attraverso il quale è possibile applicare dei metodi di apprendimento automatici (learning methods) ad un insieme di dati (dataset), e analizzarne il risultato [4]. Successivamente, l'articolo propone un ulteriore modulo dello stesso sistema finalizzato ad acquisire, in maniera diffusa e allargata sulla popolazione di interesse, una serie di informazioni di natura genetica, non ancora completamente disponibili, al fine di poterle incrociare con quelle attualmente già acquisite.

2. TECNICHE DI ESTRAZIONE DELLA CONOSCENZA ATTRAVERSO APPROCCI SUPERVISIONATI

Le strategie di data mining posso essere classificate in generale come strategie supervisionate e strategie non supervisionate. L'apprendimento supervisionato, in particolare, costruisce modelli e estrae regole tramite utilizzo di attributi di ingresso per predire valori di attributi di uscita. Ad esempio, introducendo nel sistema informazioni sullo stile di vita, sulla dieta alimentare, etc. (attributi in ingresso), si può ricavare la probabilità di contrarre una certa malattia (attributo in uscita). Una particolare tipologia di apprendimento supervisionato, con molteplici applicazioni in ambito medico e sanitario è rappresentato dalla costruzione di un albero decisionale, ovvero di una struttura molto semplice e di facile utilizzo in cui i cosiddetti nodi non terminali rappresentano uno o più attributi e i nodi terminali i risultati ottenibili. Questo paragrafo fornisce i concetti alla base di alcuni strumenti informatici disponibili all'interno del software Weka. Tali strumenti consentono di estrarre eventuali relazioni funzionali

esistenti fra variabili di un determinato scenario; ad esempio, se una certa configurazione di informazioni sintomatologiche ci offre la possibilità di effettuare una diagnosi accurata e robusta.

Premettiamo alla descrizione teorica un esempio concreto [3], per consentire al lettore di familiarizzare con la terminologia e, soprattutto, di orientarsi nella propedeutica selezione delle informazioni necessarie per utilizzare gli strumenti del software . Consideriamo un dataset (tabella 2.1) secondo il formato attributo-valore dove la variabile dipendente (Y) in questo caso è la Diagnosi, mentre i vari sintomi sono le variabili indipendenti (Xi).

X1	X2	X3	X4	X5	X6	Y
ID paziente	**Mal di gola**	**Febbre**	**Ghiandole ingrossate**	**Congestione**	**Mal di testa**	**Diagnosi**
1	**SI**	**SI**	**SI**	**SI**	**SI**	**STREPTOCOCCO**
2	**NO**	**NO**	**NO**	**SI**	**SI**	**ALLERGIA**
3	**SI**	**SI**	**NO**	**SI**	**NO**	**RAFFREDDORE**
4	**SI**	**NO**	**SI**	**NO**	**NO**	**STREPTOCOCCO**
5	**NO**	**SI**	**NO**	**SI**	**NO**	**RAFFREDDORE**
6	**NO**	**NO**	**NO**	**SI**	**NO**	**ALLERGIA**
7	**NO**	**NO**	**SI**	**NO**	**NO**	**STREPTOCOCCO**
8	**SI**	**NO**	**NO**	**SI**	**SI**	**ALLERGIA**
9	**NO**	**SI**	**NO**	**SI**	**SI**	**RAFFREDDORE**
10	**SI**	**SI**	**NO**	**SI**	**SI**	**RAFFREDDORE**

Tabella 2.1 – Dati di training in forma tabellare

Anche se il dataset è piccolo, potrebbe essere difficile per noi sviluppare una rappresentazione generale senza avere informazioni sull'importanza relativa dei singoli attributi e sulle possibili relazioni fra gli attributi. Esiste un apposito algoritmo di apprendimento supervisionato che svolge tale compito chiamato C4.5 [5]. Tale algoritmo generalizza un insieme di casi di input creando un albero decisionale.

Un albero decisionale è una struttura in grado di generare delle regole decisionali che hanno già dimostrato un sufficiente livello di funzionalità. Si basa sull'idea che è propria del metodo scientifico in generale, di risalire dal comportamento sperimentato in pochi casi noti, alla definizione di una regola utilizzabile più in generale. I casi utilizzati per costruire il citato albero sono detti dati di training.

La determinazione della efficacia del modello, ai fini di una generalizzazione dei risultati, può essere, poi, determinata applicandolo ad un insieme di dati di prova o test set. La costruzione effettiva di un albero decisionale avviene mediante l'utilizzo iniziale di un certo numero di attributi, scelti tra quelli ritenuti maggiormente in grado di differenziare i concetti che devono essere appresi.

Se l'albero decisionale classifica correttamente le altre osservazioni disponibili, la procedura termina, se invece un'osservazione non è classificata correttamente, viene aggiunta al sottoinsieme di training selezionato e viene costruito un nuovo albero.

Questo processo ricorsivo continua finché non viene generato un albero che classifichi correttamente tutte le osservazioni non selezionate.

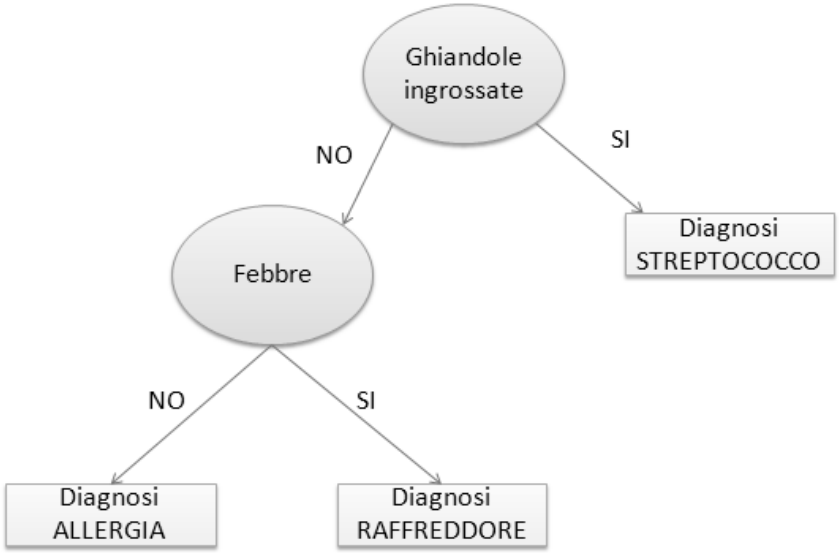

Figura 1 – Albero decisionale ottenuto dai dati della Tabella 2.1

L'albero in figura 1 è stato creato a partire dalla tabella 2.1 ed evidenzia le regole che generalizzano i dati della tabella, in particolare :
-Se un paziente ha le ghiandole ingrossate, la diagnosi è affezione da streptococco;
-Se un paziente non ha le ghiandole ingrossate e ha febbre, la diagnosi è raffreddore;
-Se un paziente non ha le ghiandole ingrossate e non ha febbre, la diagnosi è allergia.
L'albero decisionale ci offre la possibilità di diagnosticare accuratamente le malattie contenute nel dataset osservando semplicemente se un paziente ha le ghiandole ingrossate e la febbre. Gli attributi (Xi) Mal di gola, Congestione e Mal di testa non giocano un ruolo fondamentale nella determinazione della diagnosi (Y). Si può facilmente intuire che l'albero decisionale ha generalizzato i dati e ci ha offerto una sintesi degli attributi e delle loro relazioni che sono importanti per effettuare una diagnosi accurata.
Il modello ottenuto tramite i dati di training è utile perché è in grado di classificare correttamente nuovi casi la cui classificazione non è nota. Per determinare con quale efficacia il modello può essere generalizzato, possiamo provare ad applicarlo su un insieme di dati di prova o Test Set. I casi del test set hanno una classificazione nota, quindi possiamo confrontare le classificazioni dei casi del test set stabilite dal modello con quelle esatte. La correttezza delle classificazioni del test set ci fornisce un'indicazione sul comportamento futuro del modello.
Utilizziamo l'albero decisionale per classificare i seguenti casi:

X1	X2	X3	X4	X5	X6	Y
ID paziente	**Mal di gola**	**Febbre**	**Ghiandole ingrossate**	**Congestione**	**Mal di testa**	**Diagnosi**
11	**NO**	**NO**	**SI**	**SI**	**SI**	**?**
12	**SI**	**SI**	**NO**	**NO**	**SI**	**?**
13	**NO**	**NO**	**NO**	**NO**	**SI**	**?**

Tabella 2.2 – Test Set

Dal momento che il paziente con ID=11 ha le ghiandole ingrossate, seguiamo il ramo di destra dell'albero a partire dal nodo radice dell'albero decisionale. Il ramo di destra conduce ad un nodo terminale che indica che il paziente ha un'affezione da STREPTOCOCCO. Il paziente con ID=12 non ha le ghiandole ingrossate, perciò seguiamo il ramo di sinistra e valutiamo il valore dell'attributo Febbre. Dal momento che la febbre è presente, la diagnosi è RAFFREDDORE.
Lasciamo al lettore verificare che il paziente con ID=13 è affetto da ALLERGIA.
E' importante sottolineare che i modelli di apprendimento supervisionato sono ideati per classificare, stimare e/o prevedere comportamenti futuri. Per alcune applicazioni lo scopo è costruire modelli con un alto livello di accuratezza delle previsioni. Tale accuratezza può essere valutata tramite la **matrice di confusione (tabella 2.3)**.
Per valutare l'accuratezza di tale matrice basta sommare i valori contenuti nella diagonale principale e dividere il risultato per il numero totale di casi impiegati come dati di prova. Per esempio, se applichiamo il modello a 100 casi di prova e la somma dei valori lungo la diagonale principale della matrice di confusione è pari a 70, l'accuratezza del modello è pari a 70% o 0,70. Da questo valore spesso viene ricavato il tasso di errore come complemento a 1 dell'accuratezza; nell'esempio in esame, il tasso di errore è 1 - 0,70 = 0,30. Un esempio pratico di utilizzo di tale matrice viene fornito nel paragrafo 3.

Decisione calcolata			
	C_1	C_2	C_3
C_1	C_{11}	C_{12}	C_{13}
C_2	C_{21}	C_{22}	C_{23}
C_3	C_{31}	C_{32}	C_{33}

Tabella 2.3 - Matrice di confusione con tre classi

3. IL PROCEDIMENTO DI ESTRAZIONE DELLA CORRELAZIONE DELLE CAUSE DI MORTALITA' PER ALCUNI DEI TUMORI PIU' FREQUENTI QUALI AD ESEMPIO IL TUMORE AL COLON RETTO, IL TUMORE AL SENO E IL TUMORE AI POLMONI

Nel presente articolo è stata presa in considerazione la banca dati dell'Istat "Healt For All" (aggiornata al Dicembre 2008), della quale sono stati esaminati alcuni particolari indici utilizzati come attributi del nostro DataBase importato in WEKA. In particolare, fra le informazioni disponibili, sono state prese a oggetto della nostra analisi il "Luogo", il "Tasso di mortalità maschile dovuto al tumore al colon", la "Cena come pasto principale" e il "Consumo settimanale di pesce".

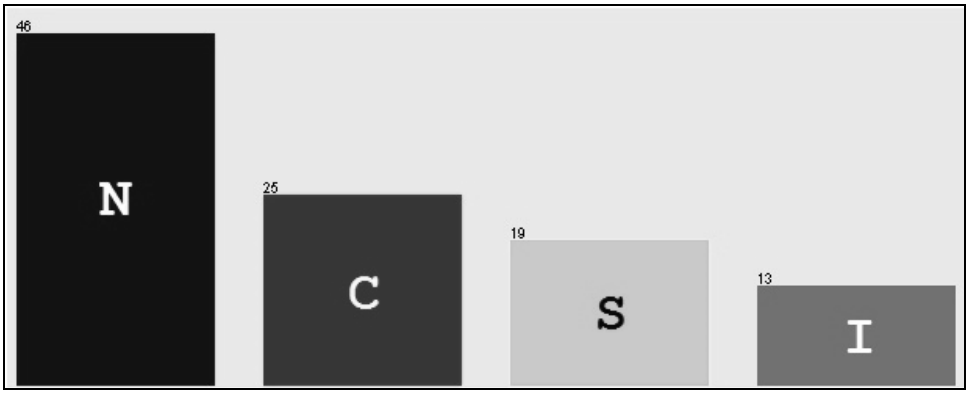

Fig. 2: La figura riporta la distribuzione delle 113 province italiane disponibili nel data base ordinate rispettivamente per Nord (46 province), Centro (25 province), Sud (19 province) e Isole (13 province).

Dopo aver importato in Weka il database, si è passati ad una fase di Pre-processing (pre-elaborazione dei dati). Tale fase è fondamentale per dare in input agli algoritmi del Weka un set di dati filtrato e organizzato nella maniera migliore per estrarre conoscenza. Un importante passo di pre-processing è la discretizzazione (conversione di un attributo numerico in nominale). A seguito della discretizzazione i valori numerici dell'attributo "Tasso di mortalità per tumore al colon" (fig.3) si sono distribuiti secondo due valori nominali : BASSO e ALTO. Si è quindi passati all'applicazione dell'algoritmo J48 per l'estrazione dell'informazione e delle regole dai dati.

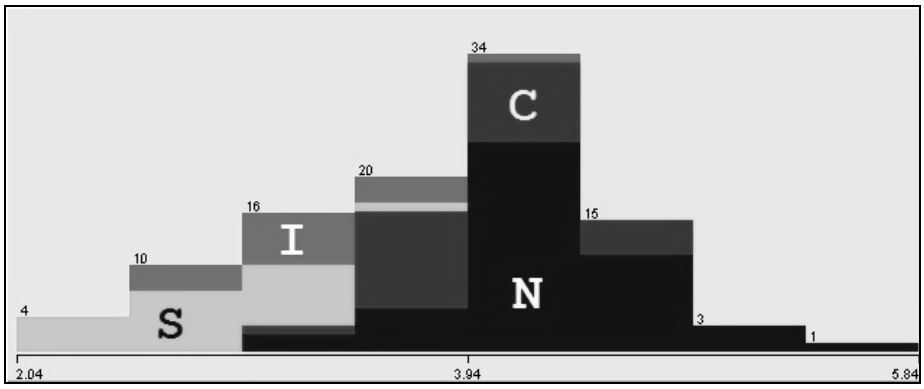

Fig. 3: Dalla figura (sulle ascisse è riportato il tasso di mortalità legato al tumore al colon maschile) si evince come la maggior parte delle province del Nord (N) e del Centro (C) Italia abbia un alto tasso di mortalità , al contrario del Sud (S) e delle Isole (I).

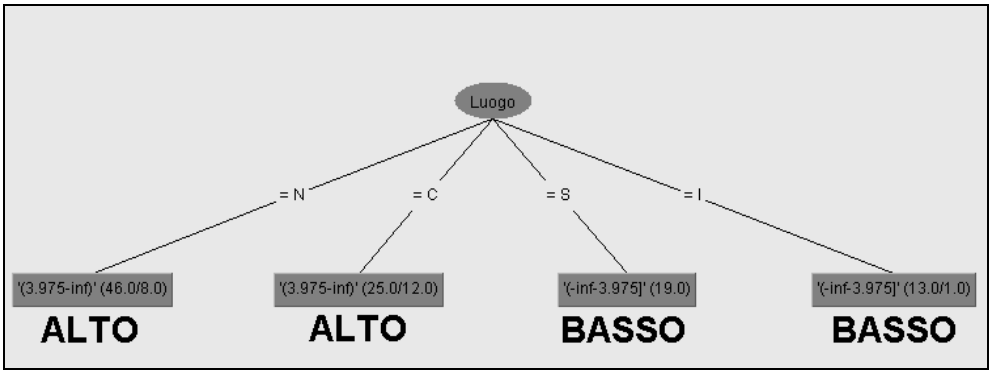

Fig. 4: Albero non binario delle regole relativo alla mortalità per tumore al colon maschile in relazione al luogo.

L'albero ottenuto tramite l'esecuzione dell'algoritmo J48 presenta la relazione (rappresentata dai rami) esistente fra il luogo (X) e il tasso di mortalità (Y) per il tumore al colon (quest'ultimo rappresentato dalle foglie dell'albero). E' evidente che valori elevati del tasso di mortalità sono presenti al NORD e al CENTRO Italia.
Nell' Italia meridionale e sulle isole sono presenti dei tassi di mortalità più bassi.
Insieme all'albero vengono presentati i relativi indici statistici che consentono di interpretare i dati estratti. Dal report si evince che le province che rispettano la logica dell'albero sono 82/103 (il 79.6%), mentre le province che non vengono classificate correttamente sono 21/103 (il 20.4%).

Dall'albero, si può notare come delle 46 province del Nord, 8 non rispettano il vincolo di classificazione trovato, del centro 12 su 25 non sono correttamente classificate e delle isole solo 1 su 13. Mentre per quanto riguarda il ramo del sud, vi è una corretta classificazione per tutte le 19 province.

Istanze Classificate Correttamente	82	79.6117%
Istanze Classificate Non Correttamente	21	20.3883%

Decisione calcolata		
	BASSO	***ALTO***
BASSO	*31*	*20*
ALTO	*1*	*50*

Tabella 3.1 - Matrice di confusione (31 province del database sono state classificate come valore BASSO TASSO DI MORTALITA' ; 50 province del database sono state classificate come valore ALTO TASSO DI MORTALITA' ; 21 province non sono state classificate correttamente)

Le province non correttamente classificate sono evincibili dalla matrice di confusione. Infatti, sulla diagonale MAGGIORE sono presenti le province classificate correttamente (31+51=82) e nelle altre zone della matrice (la diagonale secondaria) sono presenti le province non classificate correttamente (1+20=21).

E’ possibile ottenere una rappresentazione binaria (rappresentazione di più facile consultazione in cui da ciascun nodo partono solo due rami) dell’albero ricavato dal J48 impostando a VERO il valore “BINARY SPLIT” nelle opzioni del Weka.

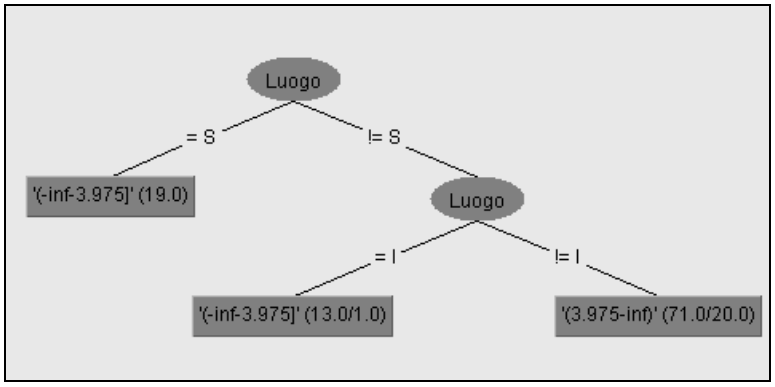

Fig. 5: Albero binario delle regole relativo alla mortalità per tumore al colon maschile in relazione al luogo.

Trascurando ora le informazioni del luogo e considerando invece un DataBase di informazioni relative alle abitudini alimentari è possibile notare l’importanza del regolare consumo di pesce all’interno delle abitudini alimentari settimanali.

Tale correlazione si può evincere studiando gli istogrammi dei relativi attributi.

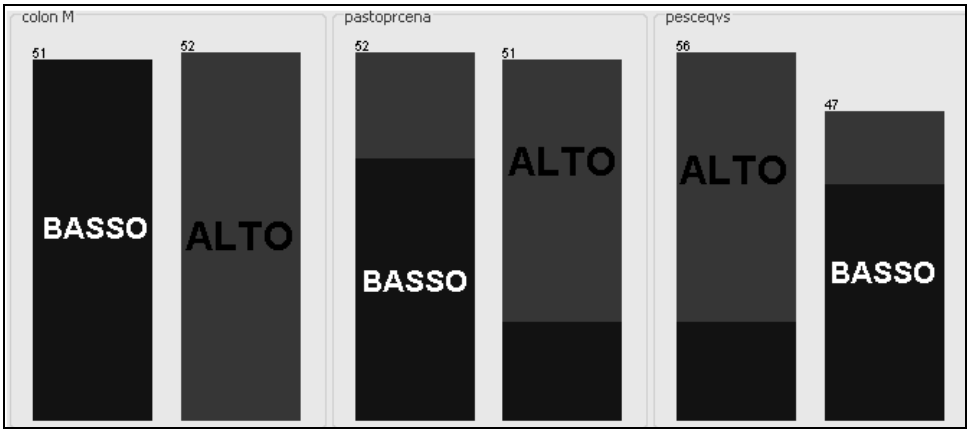

Fig. 6: Istogrammi relativi alla discretizzazione del tasso di mortalità e delle abitudini alimentari (cena come pasto principale e consumo settimanale di pesce).

Ci si aspetta che tali relazioni visibili negli istogrammi vengano riportate in forma alberata. Infatti, l’albero J48 ottenuto presenta il nesso di alta mortalità per tumore al colon legato al basso consumo di pesce alla settimana, mentre un alto consumo di pesce alla settimana unito ad un’alimentazione corretta basata sul pranzo come pasto principale è correlato ad un basso tasso di mortalità. Al contrario, un’alimentazione basata sulla cena come pasto principale anche in unione ad un alto consumo di pesce porta ad un alto tasso di mortalità.

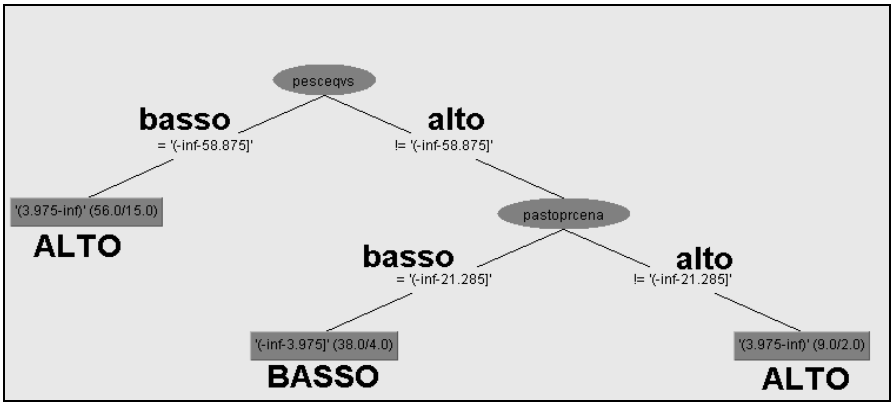

Fig. 7: Albero binario delle regole relativo alla mortalità per tumore al colon maschile in relazione alle abitudini alimentari (consumo di pesce e cena come pasto principale).

Un'ulteriore analisi può essere realizzata considerando un database dove (Y) risulta essere il tumore al seno. In questo caso si nota il nesso fra il tumore al seno femminile e l'età media del primo parto. Infatti, uno dei fattori di rischio legati al BREAST CANCER è proprio la tarda età del concepimento del primo figlio. Lanciando il J48 si può notare come un'età del primo concepimento inferiore ai 32 anni implichi un rischio inferiore di contrarre il tumore al seno, cosa che invece è molto più probabile superata tale soglia.

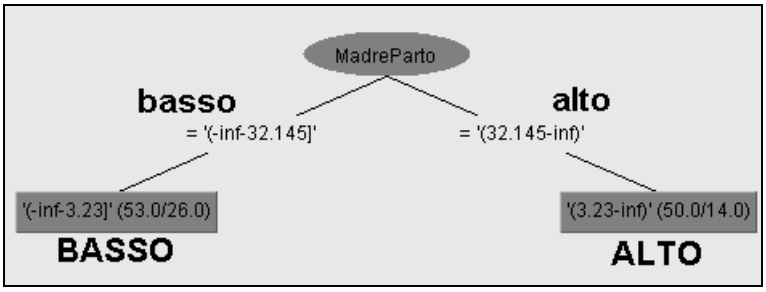

Fig. 8: Albero binario delle regole relativo alla mortalità per tumore al seno in relazione alle età della paziente al primo parto.

Considerando,invece, un database dove (Y) è la patologia "tumore ai polmoni" , si nota come sia meno netto il legame con il luogo, anche se analizzando l'istogramma relativo vi è una prevalenza per il nord Italia. Tuttavia, tale patologia è legata all'inquinamento dell'aria, presente maggiormente in zone ad alto sviluppo industriale quale il nord Italia. L'istogramma che correla il tasso di mortalità all'inquinamento mostra un netto legame fra questi due fattori. Lanciando l'algoritmo J48 si ottiene il relativo albero esplicativo. Da tale albero si nota come l'inquinamento basso è indice di un basso tasso di mortalità legato a tale patologia tumorale. Di contro, un alto inquinamento dell'aria è indice di un alto tasso di mortalità.

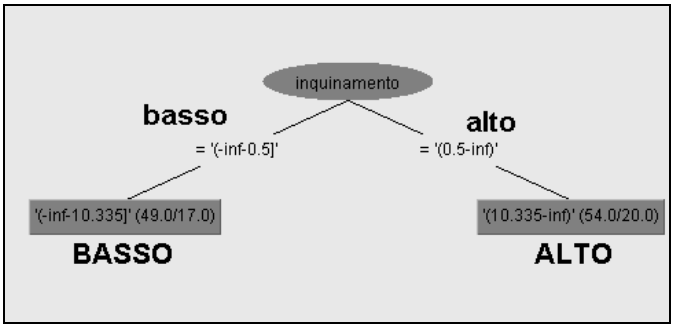

Fig.9: Albero binario delle regole relativo al tasso di mortalità per tumore ai polmoni in relazione al tasso di inquinamento dell'ambiente in cui si vive.

4. UN INNOVATIVO SISTEMA DI ACQUISIZIONE E MEMORIZZAZIONE DEI DATI CLINICI PER LO SCREENING RELATIVO ALLA FAMILIARITA' DEL CANCRO AL SENO

Nel presente paragrafo si descrive in maniera sintetica il prototipo del sistema web-based sviluppato dagli autori Magarelli, Santarcangelo e Bevilacqua con il quale si intende nel prossimo futuro fornire uno strumento di acquisizione di dati clinici di pazienti al fine di poter pervenire alla creazione di una completa base di dati su cui poter operare indagini di screening basate sulle tecniche descritte nel paragrafo **3** . Il presente prototipo, attraverso un opportuno sistema di pagine web, progettato per fornire una interfaccia di inserimento dei dati ritenuti di interesse quali fattori di rischio per una determinata malattia, consente una successiva interrogazione degli stessi attraverso una base di conoscenza realizzata sullo schema del thesaurus MultiWordNet. A titolo esemplificativo, nella figura 10 viene mostrato uno screenshot del form per l'inserimento di comuni dati anagrafici, nonché di dati quali il codice fiscale dei genitori, particolarmente utile per implementare criteri che tengano conto di protocolli relativi alle indagini di familiarità in tumori quali per esempio quello al colon retto ad incidenza prettamente maschile e quello al seno.

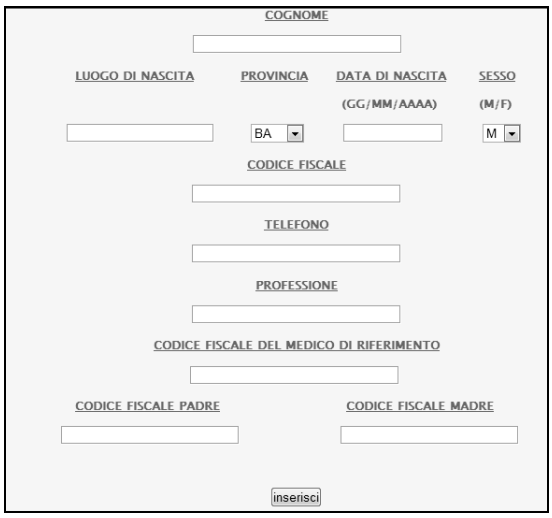

Fig. 10: Form di acquisizione di dati anagrafici e familiari.

In particolare il sistema consente, attraverso l'acquisizione delle informazioni relative ai codici fiscali dei genitori del paziente, di poter mappare lo stesso nell'ambito delle relazioni di familiarità. Dal momento che tutti i campi sono correlati tra loro mediante particolari relazioni (triple soggetto-predicato-oggetto come quelle riportate in Tab. 4.1) sarà possibile risalire alle informazioni genetiche relative per esempio ai fratelli, alle sorelle, ai cugini, ai nonni e agli zii oltre che ovviamente a quelle dei genitori.

SIMBOLO	RELAZIONE	SIGNIFICATO
^p	Meronimia	Ha parte in
h	Professione	Ha lavoro come
e	Effettuare	Ha effettuato
^s	Possibile determinazione	Può determinare
u	Possedere	Possiede
c	Conseguenza	E' conseguenza di
m	Ha madre	Ha come madre
p	Ha padre	Ha come padre
d	Ha medico	Ha come medico

Tab. 4.1: Tabella delle relazioni presenti all'interno del database e codificate nel sistema.

Dall'esame del form presente in fig. 12, relativo alla patologia del tumore al seno, è possibile notare come, grazie alle relazioni implementate circa la familiarità , sia possibile effettuare delle analisi inferenziali basate sul recupero dei dati genomici dei parenti di una determinata paziente. In questo modo, il prototipo realizzato si propone come uno strumento utile per poter implementare i moderni protocolli di screening attualmente basati non solo sull'evidenza di esami diagnostici e morfologici (TC, RMN, Mammografia, ecc.) e sierologici (marcatori tumorali) ma anche sull'analisi delle mutazioni di due particolari geni chiamati BRCA1 e BRCA2. E' altresì evidente come attraverso la stessa interfaccia sia possibile, per esempio, nel caso del tumore al colon retto acquisire, informazioni specifiche di natura alimentare quali ad esempio il *consumo di pesce una volta a settimana e una costante attività fisica* (**fattore positivo**), o *l'alto consumo di carni rosse* o di alcool (**fattore negativo**), in aggiunta a tutte le *informazioni correlate ai geni* che intervengono nell'insorgenza di questo tipo di malattia multifattoriale, e nel caso del tumore ai polmoni informazioni sulla *quantità di sigarette fumate al giorno*, piuttosto che informazioni correlate ad eventuali esposizioni derivanti dall'attività lavorativa. Al tempo stesso, il sistema proposto, sarà in grado di fornire, sulla base delle regole che potranno essere estratte e ponderate con opportuni pesi (attraverso le tecniche inferenziali presentate al paragrafo **2**), il calcolo di indici di rischio multifattoriali particolarmente utili, a giudizio degli autori, per supportare ed indirizzare politiche di organizzazione sanitaria all'interno, per esempio, di particolari regioni se non addirittura di distretti sanitari.

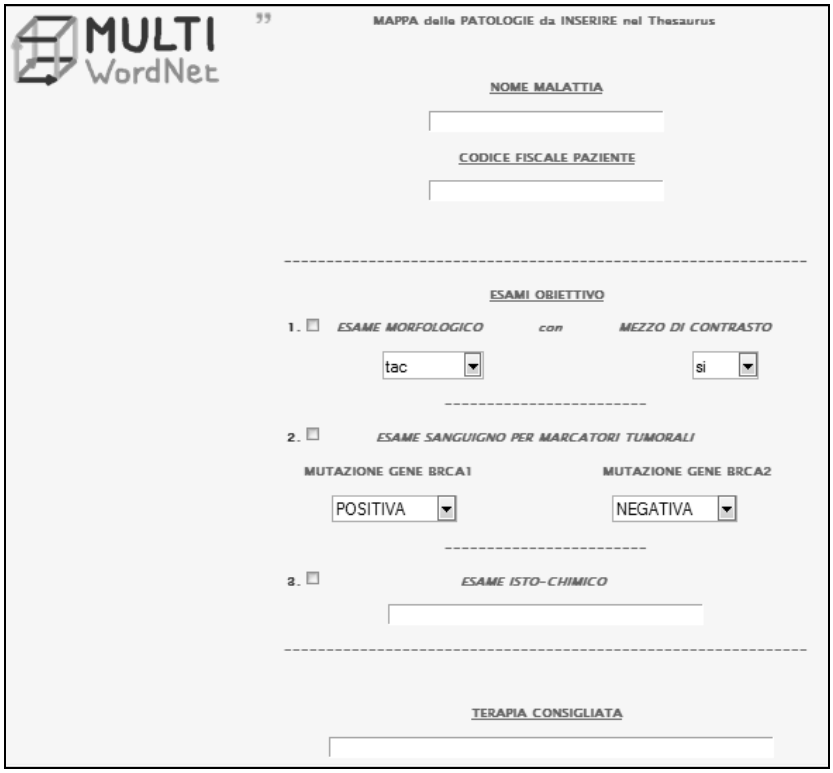

Fig. 11: Form per l'inserimento dei dati sul tipo di patologia, gli esami obiettivo effettuati e la terapia consigliata dal medico. Esso è contestualizzato al test genetico per il tumore al seno.

Fig. 12: Schermata della pagina di ricerca all'interno del thesaurus.

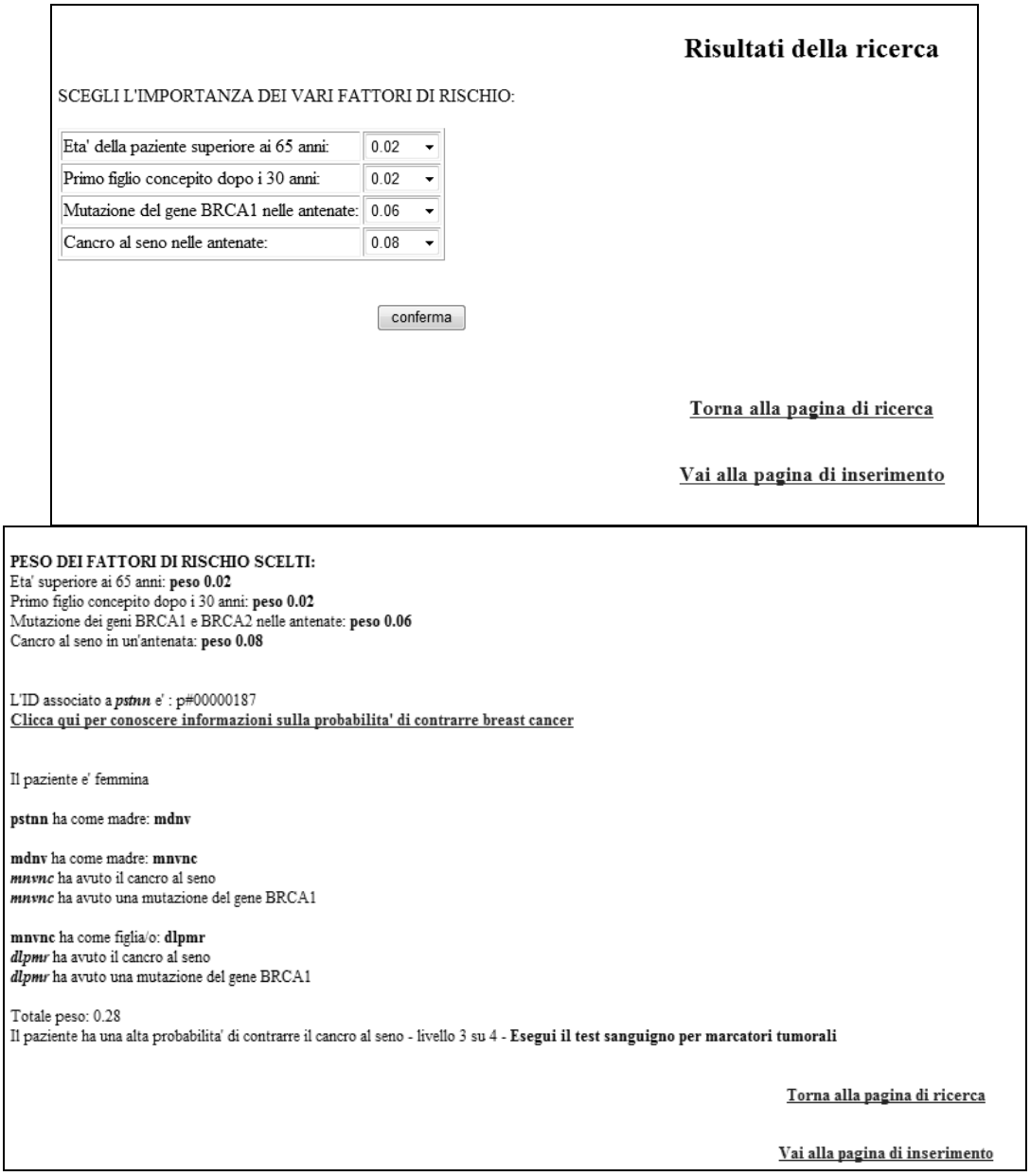

Fig. 13 : Modulo per il calcolo dell'indice di rischio relativo al tumore al seno, basato su dati quali l'età della paziente, l'età di concepimento del primo figlio, eventuali casi di stessa malattia in familiari, con la possibilità al momento di variare i pesi di ciascun fattore di rischio.

5. CONCLUSIONI E SVILUPPI FUTURI

Le conclusioni che si possono trarre ci portano a dire che il Thesaurus possa costituire un'interessante applicazione futura come base di conoscenza (KB) per supportare grazie agli strumenti forniti anche da WEKA una serie di scelte a supporto di processi decisionali non soltanto in ambito medico, ma soprattutto in ambito sanitario e di politica sanitaria. Il progetto che è stato appena presentato rappresenta la base da cui partire per sviluppare nuove applicazioni, che possano rendere assolutamente ampio il

suo ambito d'utilizzo. Soprattutto può essere integrato in un sistema di ricerca semantica per dati e cartelle cliniche. Questo thesaurus, infatti, opportunamente popolato, costituirà la **KB** (*Knowledge Base*) da cui estrarre le informazioni d'interesse in uno scenario che potrebbe essere appunto quello rappresentato nella Fig. 14.

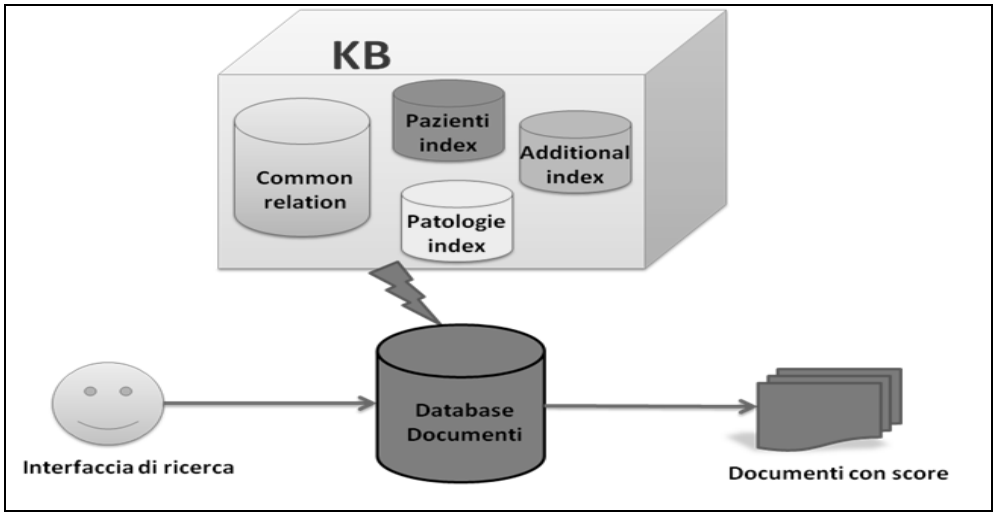

Fig. 14: Motore di ricerca basato sulla Knowledge Base (KB) costituita dal thesaurus.

BIBLIOGRAFIA

[1] **Bevilacqua V. et al**. (2007) Data mining techniques in a CGH-based breast cancer subtype profiling: the biological perspective – Proceedings of IEEE - CIBCB pp. 9-16

[2] **Bevilacqua V. et al.** (2009) A New Ontological Probabilistic Approach to the Breast Cancer Problem in Semantic Medicine. ICIC (2) 2010: pp. 59-68

[3] **Roiger R.J., Geatz M.W.** 2004 Introduzione al Data Mining – McGraw-Hill

[4] *http://it.wikipedia.org/wiki/Weka*

[5] **Quinlan J.R.** (1993) Programs for machine learning. San Mateo CA: Morgan Kaufmann