UNIVERSITA' DEGLI STUDI DI PADOVA

FACOLTA' DI SCIENZE STATISTICHE

CORSO DI LAUREA SPECIALISTICA
IN STATISTICA ED INFORMATICA

RAFFRONTO FRA METODI DI APPRENDIMENTO DI RETI BAYESIANE SU UN INSIEME DI DATI REALI

Relatore: Ch.ma Prof.ssa Adriana Brogini

Correlatore: Ch.ma Prof.ssa Nadia Minicuci

Laureanda: Federica Calabretti

ANNO ACCADEMICO: 2007/2008

Alla mia splendida famiglia

INDICE

1 INTRODUZIONE

L'umanità è abituata a vivere in condizioni di incertezza: il mondo è complesso e non facilmente prevedibile. Le nostre credenze, le nostre decisioni si basano su molti fattori, quali l'esperienza e la conoscenza.

Ogni analisi statistica si prefigge l'obbiettivo di spiegare, attraverso l'analisi dei dati, il comportamento di uno o più aspetti dello studio in esame; ciò può essere fatto esplorando le relazioni esistenti tra le variabili in esame.

In questa tesi per individuare le relazioni tra variabili si farà ricorso ai network (o rete) bayesiani: una **rete bayesiana** è la rappresentazione grafica di un modello probabilistico, ovvero la riproduzione di una distribuzione di probabilità su un insieme di variabili X .

Questo approccio unisce la metodologia statistica e l'intelligenza artificiale. Tale binomio è uno strumento vantaggioso sotto molteplici aspetti, quali la possibilità di simulare e replicare situazioni anche molto elaborate, eseguire algoritmi complessi in tempi di gran lunga inferiori rispetto allo svolgimento degli stessi senza l'ausilio di un calcolatore, immagazzinare molte informazioni.

I network bayesiani sono in grado di mettere in evidenza la struttura di un fenomeno mediante una rappresentazione grafica intuitiva, permettendo anche ai non "esperti" del settore di comprendere le relazioni di in/dipendenza; mediante questo approccio è possibile apprendere informazioni dai dati ed al contempo introdurre nell'analisi il giudizio di un esperto del settore (medico in studi clinici, etc.).

Obbiettivo di questa tesi è confrontare le strutture che derivano da due differenti metodi di apprendimento della struttura di una rete bayesiana utilizzando un insieme di dati reali.

Saranno affrontate esclusivamente variabili aleatorie discrete[1] e dataset completi[2]: la maggior parte dei metodi presenti in letteratura richiedono, infatti, dataset in cui non sono presenti dati mancanti e variabili discrete. Le stesse condizioni sono necessarie anche per i software utilizzati per il learning della struttura e de i parametri delle rete bayesiane.

La tesi è così strutturata:

- Capitolo 2: "Network bayesiani".

  Definizione di una rete bayesiana ed proprietà che la caratterizzano; sono trattate inoltre le relazioni di in/dipendenza in un grafo.

[1] In presenza di variabili continue, occorre discrettizzarle.

[2] In cui non sono presenti dati mancanti.

- Capitolo 3: “Inferenza nei network bayesiani”.

  Analisi della fase d’inferenza di una rete bayesiana ed rassegna dei principali algoritmi per l’inferenza esatta ed approssimata.

- Capitolo 4: “Apprendimento dei network bayesiani”.

  Lo scopo di questa sezione è approfondire la fase del learning (o apprendimento) sia dei parametri che della struttura di una rete bayesiana; in particolare l’attenzione sarà rivolta a quest’ultima specificando i metodi e gli algoritmi presenti in letteratura.

  L’algoritmo K2 e l’algoritmo BNPC sono trattati in modo approfondito.

- Capitolo 5: “Analisi di un insieme di dati reali”.

  Descrizione del problema, delle variabili e delle unità statistiche. Inoltre è presente una panoramica dei due software utilizzati per le analisi.

- Capitolo 6: “Conclusioni”.

2 NETWORK BAYESIANI

Le reti bayesiane sono definite tramite la specificazione di due componenti:

a) la *componente qualitativa*: un grafo diretto aciclico (DAG), indicato con *G=(V,A)* detto struttura della rete:
   - ✓ *V* sono i nodi[3] che sono in corrispondenza biunivoca con l'insieme *X* di variabili aleatorie;
   - ✓ gli archi *A* sono coppie ordinate di elementi di *V* .

   Ogni arco denotato con $X_i \rightarrow X_j$ rappresenta la dipendenza condizionata esistente tra i due nodi; i genitori di un nodo X_i sono denotati da $Pa(X_i)$ i figli con $Ch(X_i)$.

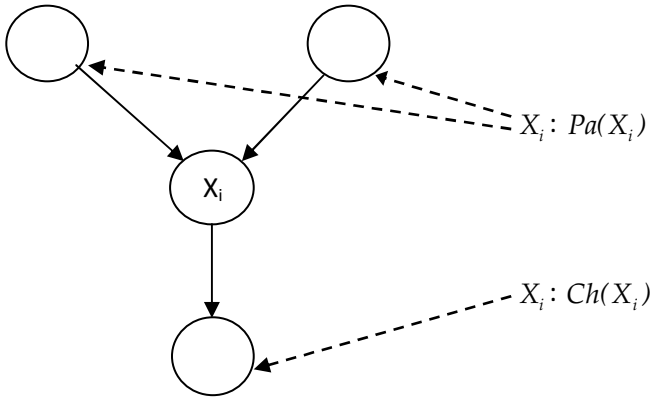


FIGURA 1: ESEMPIO RETE BAYESIANA

b) la *componente quantitativa:* un insieme di distribuzioni locali di probabilità, ciascuna associata ad una variabile e condizionata ad ogni configurazione dei suoi genitori. L'insieme delle distribuzioni locali di probabilità specificano la distribuzione congiunta dell'insieme di variabili. Queste distribuzioni di probabilità sono dette anche **parametri** delle reti bayesiane.

Esse sono contraddistinte da alcune caratteristiche:

- ✓ Esprimono legami non deterministici tra le variabili aleatorie;
- ✓ Le conoscenze preliminari sul sistema permettono, almeno parzialmente, di determinare la struttura della rete;
- ✓ La distribuzione di ogni variabile è influenzata solamente dalle distribuzioni dei suoi diretti vicini all'interno della struttura. Quindi un nodo ha una tabella di probabilità condizionata che quantifica

[3] In queste tesi verrà utilizzato il termine nodo e variabile in modo interscambiabile.

gli effetti che i genitori hanno sul nodo. Un nodo che non ha genitori contiene una tabella di probabilità marginale;

- ✓ Se il nodo è discreto contiene una distribuzione di probabilità sugli stati della variabile che rappresenta;
- ✓ Il grafo non ha cicli diretti.

L'utilizzo delle reti bayesiane ha dei vantaggi: la rappresentazione grafica e la struttura delle relazioni tra variabili aleatorie risulta intuitiva e di facile comprensione; sono utilizzabili anche per insiemi di dati incompleti; facilitano le combinazioni del dominio di conoscenza dei dati, permettendo la possibilità di specificare dei giudizi soggettivi di esperti sul modello.

2.1 Le relazioni di in/dipendenza

La **in/dipendenza** è intesa come uno strumento che identifica le strutture di relazione tra le variabili aleatorie in esame e focalizza l'attenzione su ciò che è *rilevante* per studiare un fenomeno.

Per poter descrivere interamente una situazione bisogna tener conto di altre due componenti: il *condizionamento* e la *verosimiglianza*. Il condizionamento serve per mettere in evidenza come il comportamento di una variabile può modificare l'andamento di un'altra, la verosimiglianza per poter identificare quali situazioni sono induttivamente più probabili di altre. L'esempio che segue ha lo scopo di chiarire questi concetti:

1. Holmes e Watson sono vicini di casa;
2. Una mattina Holmes andando al lavoro *nota* (apprende) che il suo prato è bagnato e si chiede se ha lasciato l'impianto di irrigazione acceso o se ha piovuto;
3. Guardando il giardino di Watson *nota* (apprende) che anche questo è bagnato;
4. Holmes pensa: "poiché anche il giardino del vicino è bagnato, probabilmente questa notte ha piovuto";
5. Pensa inoltre: "la pioggia spiega come mai il mio prato è bagnato, e quindi non c'è ragione di ritenere che l'impianto di irrigazione sia stato acceso.

Quanto appena descritto è la trascrizione di una situazione che può sembrare banale quanto consueta. Si tratta ora di convertire ogni punto riportato in una struttura di relazioni di in/dipendenza.

La situazione iniziale si può così rappresentare:

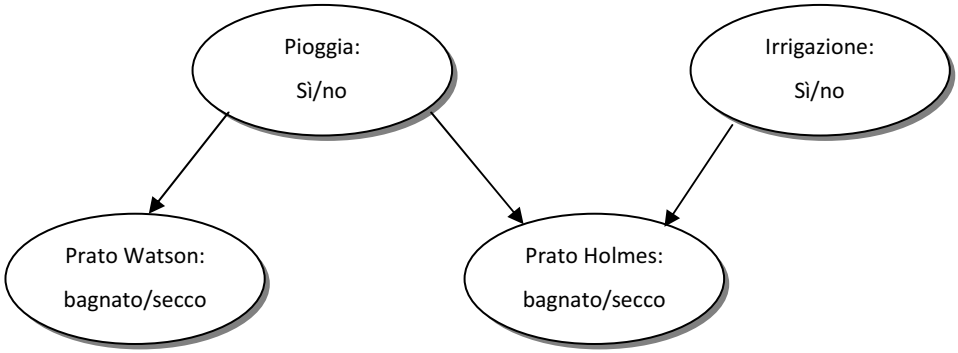


FIGURA 2: RELAZIONI DI IN/DIPENDENZA

→ Holmes apprende che il suo prato è bagnato e questo genera una propagazione dell'informazione:

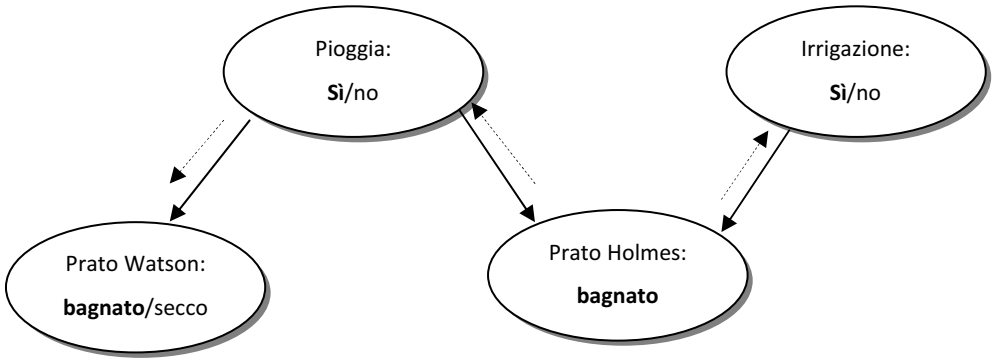


FIGURA 3: RELAZIONI DI IN/DIPENDENZA

Il fatto che il prato di Holmes sia bagnato può derivare da due situazioni: ha piovuto oppure l'impianto d'irrigazione è rimasto acceso.

→ Per poter individuare la causa Holmes osserva il prato del vicino e nota che è bagnato:

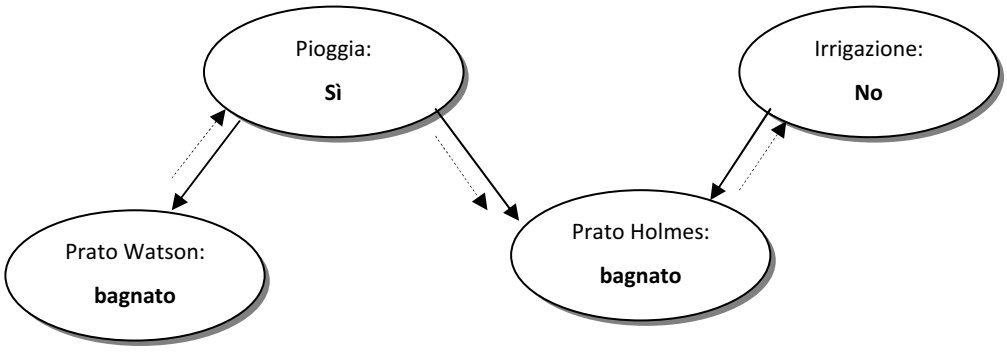


FIGURA 4: RELAZIONI DI IN/DIPENDENZA

→ Sapendo che anche il prato di Watson è bagnato, è più verosimile che abbia piovuto.

Il tipo di evidenza disponibile può influenzare la dipendenza o indipendenza tra due variabili aleatorie: la presenza di pioggia e l'aver lasciato l'impianto di irrigazione acceso sono solitamente quantità non

correlate, ma l'evidenza che il prato sia bagnato, le rende dipendenti perché l'assenza di pioggia rende più verosimile il fatto che l'impianto sia rimasto acceso.

Come si può notare sono stati utilizzati tutti e quattro i concetti introdotti per arrivare a definire le relazioni di in/dipendenza, in particolare partendo dalla conoscenza dedotta dai fatti e dall'esperienza dell'analista (in questo caso rappresentata dalla figura di Holmes) si è arrivati a costruire la struttura a cui corrisponde una probabilità, seppur qualitativa, più alta.

Il tipo di **connessione** che unisce i nodi di una rete bayesiana determina le in/dipendenze condizionate che sussistono tra le variabili aleatorie delle reti ed il passaggio di **informazione** che si attiva fra esse, inteso come influenza che la conoscenza sullo stato di una variabile può esercitare su un'altra variabile.

Definizione indipendenza:

> *date due variabili aleatorie X e Y si dice che esse sono indipendenti se la probabilità congiunta, può essere fattorizzata : $P(X,Y)=P(X)P(Y)$.*

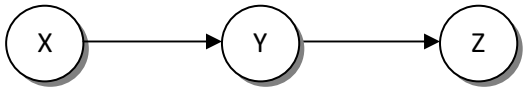
Definizione variabili aleatorie condizionatamente indipendenti:

> *Siano X , Y e Z variabili aleatorie casuali o sottoinsiemi di variabili aleatorie casuali **X**, **Y**, **Z**. X e Y sono condizionatamente indipendenti dato Z se*
>
> $$P(X=x|Y=y,Z=z)=P(X=x|Z=z)$$
>
> *per ogni valore x, y e z che le variabili aleatorie possono assumere.*

L'indipendenza condizionata sarà indicata: $X \perp Y \mid Z$

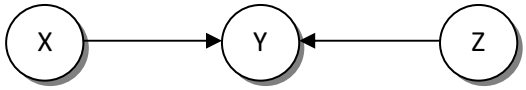
Si può affermare che una rete bayesiana consiste in un insieme di affermazioni di in/dipendenze condizionate che sono implicate dalla struttura.

1. *Connessione seriale*:



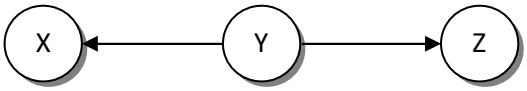
La conoscenza sullo stato della variabile X influenza la conoscenza su Y che a sua volta influenza Z . Un'informazione sullo stato di Z può influenzare la conoscenza su X attraverso Y ; se lo stato di Y è noto, allora il passaggio dell'informazione è bloccato e X e Z diventano indipendenti condizionatamente ad Y .

2. *Connessione convergente*:



Se non si hanno informazioni sullo stato di Y, eccetto quello che può essere dedotto dalla conoscenza sugli stati dei suoi genitori, X e Z risulteranno essere indipendenti e nessuna conoscenza sullo stato di uno di essi influenzerà lo conoscenza sullo stato dell'altro. Se invece si ha informazione sullo stato di Y o di qualcuno tra i suoi figli, X e Z diventano dipendenti.

3. *Connessione divergente*:



Se si conosce lo stato assunto da Y, non si ha passaggio di informazione fra i suoi figli X e Z, che risulteranno essere indipendenti.

2.2 CARATTERISTICHE E PROPRIETÀ DI UNA RETE BAYESIANA

Per quanto detto, si può definire una rete bayesiana come la rappresentazione della probabilità congiunta P sull'insieme di variabili aleatorie **X**. Utilizzando la "Condizione di Markov" si può affermare che la distribuzione è formalizzata come prodotto di un insieme di probabilità locali:

Definizione condizione di Markov:

*la probabilità congiunta P(**X**) soddisfa la condizione di Markov per un DAG G, se ogni variabile X_i è condizionatamente indipendente da ogni altra variabile escludendo i figli e i genitori, dati i genitori:*

$$X_i \perp \mathbf{X} \setminus \{Pa_i \cup Dc_i\} \mid Pa_i \qquad \forall X_i \in \mathbf{X}$$

Definizione condizione di Markov Blanket:

dato un nodo X di un network bayesiano, il Marov Blanket di un nodo X è definito come un insieme di nodi costituito dai genitori di X, dai suoi figli e dai genitori di ogni figlio del nodo X. Dalla condizione di Markov deriva che condizionandosi ad ogni nodo appartenente al Markov Blanket si ottiene che X è indipendente da tutti gli altri nodi della rete escludendo quelli appartenenti al Markov Blanket.

Per chiarire il concetto si riporta un esempio. I nodi evidenziati fanno parte del Markov Blanket:

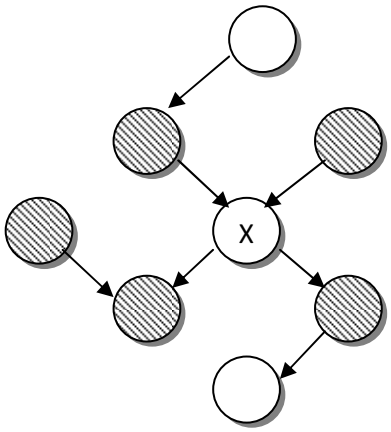


FIGURA 5: MARKOV BLANKET

Per quanto detto prima è possibile scrivere che:

$$P(X_1,\ldots,X_n)=\prod_{i=1}^{n}P(X_i \mid X \setminus X_i) \tag{2.1}$$

Per la condizione di Markov è possibile semplificare la formula precedente, ottenendo:

$$P(X_1,\ldots,X_n)=\prod_{i=1}^{n}P(X_i \mid Pa_i^G) \tag{2.2}$$

Si elencano di seguito le proprietà fondamentali di un network bayesiano:

- ✓ Condizione minimale:

Un DAG G e la probabilità associata $P(\mathbf{X})$ soddisfa la condizione minimale, se ogni sottografo $H \subset G$ non soddisfa la condizione di Markov.

Un DAG che soddisfa la condizione di Markov e la condizione minimale rappresenta la distribuzione di probabilità del grafo.

- ✓ D-separazione:

Siano i nodi X, Y e Z tre insiemi disgiunti di variabili aleatorie in un grafo orientato aciclico; X e Y si dicono d-separati da Z, ($X \perp Y \mid Z$), se non esiste un cammino tra X e Y tale che (2.1) per ogni nodo con archi convergenti appartiene a Z o ha un discendente che appartiene a Z, oppure (2.2) per qualsiasi altro nodo che non appartiene a Z. Se X e Y non sono d-separati allora sono d-connessi. Questa proprietà è considerata regola principale per l'inferenza, come dimostrato in Geiger et al. [1988b]. Geiger

et al. [1990] che mostrano che la D-separazione come regola per l'inferenza è atomic-complete[4] per le distribuzioni multinomiali e multivariate.

- ✓ Faithfulness:

Un network bayesiano composto da variabili aleatorie discrete è **faithful** se e solo se le relazioni di indipendenza condizionata deducibili dalla rete sono esattamente quelle desumibili dalla fattorizzazione della distribuzione di probabilità del grafo stesso.

In un articolo pubblicato da Meek[5], egli dimostra che le distribuzioni multinomiali faithful esistono sempre per ogni grafo aciclico diretto e per un insieme di variabili aleatorie discrete. Alcuni teoremi presenti in tale articolo sono di fondamentale importanza:

Teorema (Existence):

Per ogni grafo aciclico diretto G esiste una distribuzione P che è faithful per G.

Teorema (Measure zero):

In accordo con la misura di Lebesgue[6] su π_G^D, l'insieme di distribuzioni che sono unfaithful per G ha misura zero.

Dove π_G^D rappresenta l'insieme dei parametri indipendenti necessari per parametrizzare una distribuzione discreta per il grafo G.

Le condizioni e le proprietà introdotte sono sufficienti per specificare le relazioni tra il network bayesiano e la probabilità congiunta delle variabili aleatorie in esame. Esiste inoltre un criterio che permette graficamente di definire le relazioni tra le variabili aleatorie usando, in particolare, la condizione di Markov.

Dalla proprietà della D-separazione si ricava il seguente teorema:

Teorema (Verma e Pearl, 1988):

Se i nodi X e Y sono d-separati da Z allora X e Y sono condizionatamente indipendenti dato Z.

Sfruttando il risultato è possibile limitare i calcoli relativi alle probabilità delle variabili aleatorie.

Nel caso della connessione seriale e divergente prima introdotte, i nodi X e Y sono d-separati Y per la condizione (2), mentre nella connessione convergente X e Y sono d-separati solo se Y non è istanziata, ossia se il suo stato non è noto. In questo caso la presenza di evidenza sullo stato di Y blocca il flusso di informazione tra X e Z rendendoli indipendenti.

[4] Definizione in appendice.

[5] Riferimenti precisi in bibliografia

[6] Definizione in appendice.

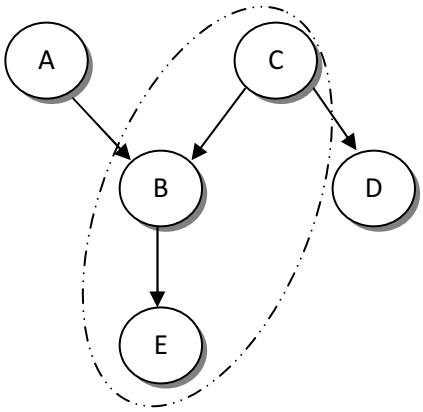


FIGURA 6: D-SEPARAZIONE

La figura mostra un esempio di d-separazione: A e D sono d-connessi dato E, poiché il cammino non diretto da A a D ha solo una connessione convergente su B, che è il genitore di E. Risultano d-separati dato l'insieme $\{C, E\}$ perché C è il nodo che collega il cammino non diretto tra A e D.

3 INFERENZA NEI NETWORK BAYESIANI

La costruzione di un modello ed il consecutivo utilizzo è l'obbiettivo principale di un'analisi statistica. Il grafo ha lo scopo di caratterizzare le in/dipendenze condizionate delle variabili aleatorie aleatorie in esame.

Il processo di costruzione di una rete viene definito **apprendimento** o **learning**: in questa fase si ricerca la struttura e le probabilità associate alla rete.

La **inferenza probabilistica** costituisce la successiva fase al processo di apprendimento della rete bayesiana.

Uno dei principali obbiettivi dell'utilizzo dei network bayesiani riguarda la possibilità di determinare la probabilità $P(X_k|e)$: ossia la probabilità a posteriori del nodo X_k data l'informazione e.

L'informazione che si ha a disposizione viene chiamata **evidenza**: la propagazione di essa consiste nell'aggiornare le distribuzioni di probabilità delle variabili aleatorie in accordo con la nuova informazione disponibile.

L'inferenza probabilistica permette di utilizzare l'informazione a disposizione e calcolare le relative probabilità; tale procedimento è NP-hard[7] in caso di molte dipendenze tra variabili aleatorie.

La propagazione dell'evidenza può avvenire dall'alto verso il basso, ossia può passare dai genitori ai discendenti, in questo caso l'obbiettivo è calcolare le probabilità dei nodi-figli dopo il passaggio dell'informazione; oppure essa può avvenire atttraverso il processo inverso, l'evidenza si ha nei discendenti e passa ai genitori, chiaramente l'attenzione probabilistica è rivolta ai nodi-genitori.

Esistono due tipi di evidenza: se lo stato assunto da una o più variabili aleatorie è noto si parla di *evidenza hard;* se invece non si conoscono con certezza i valori assunti da una o più variabili aleatorie, ma si possono fare delle affermazioni sul loro stato si parla di *evidenza soft*.

La propagazione dell'evidenza avviene applicando il teorema di Bayes:

$$P(X_1|X_2)=\frac{P(X_2|X_1)P(X_1)}{\sum_i P(X_2|X_i=x_i)P(X_1=x_i)} \tag{3.1}$$

Questa tecnica risulta efficace se il numero di variabili aleatorie non è elevato e quando le modalità che ogni variabile può assumere non sono molte; in caso contrario il calcolo richiesto è troppo oneroso.

Si prenda in considerazione l'insieme delle variabili aleatorie O e un insieme di nodi Q, le probabilità marginali sono così calcolate:

[7] NP-hard: (nondeterministic polynomial-time hard), in computational complexity theory, is a class of problems informally "at least as hard as the hardest problems in NP."

$$P(Q,O)=\sum_{X\setminus\{O\cup Q\}} P(X) \qquad P(O)=\sum_{Q} P(Q,O) \tag{3.2}$$

Inserendo l'evidenza si possono calcolare le seguenti probabilità condizionate:

$$P(Q \mid O=o)=\frac{P(Q,O=o)}{P(O=o)} \tag{3.3}$$

Tenendo in considerazione la fattorizzazione delle probabilità congiunte della rete bayesiana, il problema viene riformulato nel seguente modo:

$$P(Q|O=o)=\sum_{X\setminus\{O\cup Q\}} P(x|O=o)=\sum_{X\setminus\{O\cup Q\}}\prod_{i=1}^{N} P(X_i|Pa_i,O=o) \tag{3.4}$$

E' chiaro che questo modo di procedere porta ad un numero elevato di fattorizzazioni che cresce all'aumentare del numero di nodi. Si consideri infatti il seguente esempio:

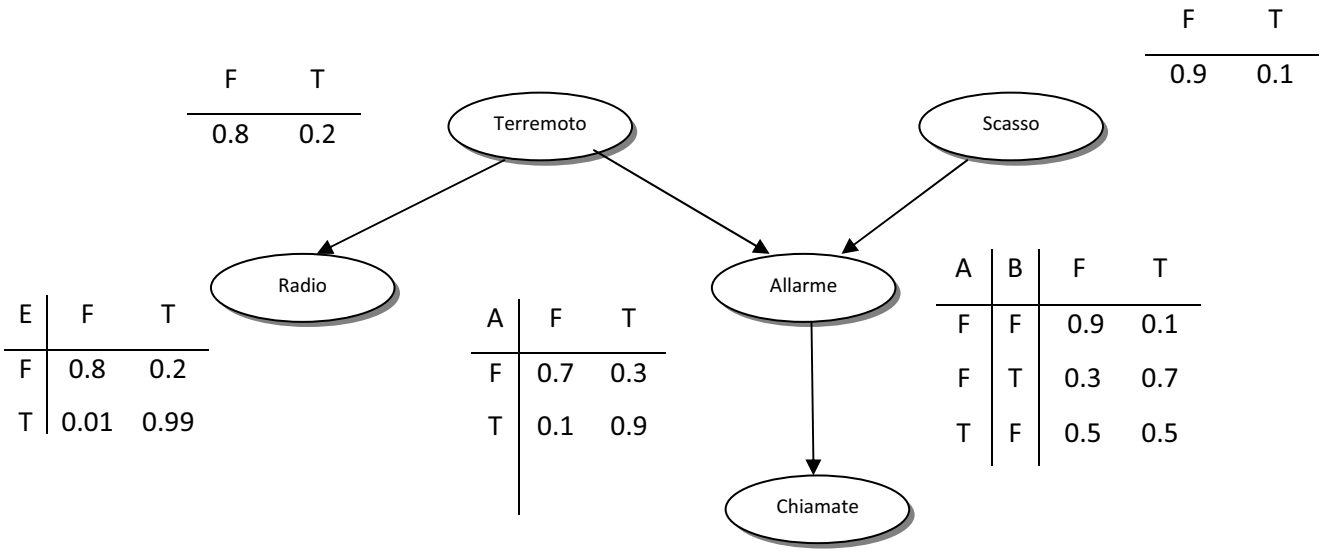


FIGURA 7: INFERENZA

Il grafo descrive la seguente situazione:

- ✓ Una chiamata alla stazione di polizia arriva in seguito all'attivazione di un allarme;
- ✓ L'allarme di un'abitazione può attivarsi per uno scasso o a causa di un terremoto;
- ✓ Il terremoto attiva generalmente un allarme di un'abitazione e la notizia della sua presenza viene trasmessa alla radio.

Si vuole calcolare la probabilità che Watson telefoni alla stazione della polizia se uno scasso è avvenuto. In altre parole si vuole calcolare la probabilità che se una chiamata arriva alla polizia, questa sia dovuta ad uno scasso e non generata dalle scosse di un terremoto.

Per alleggerire la notazione, in seguito la "chiamata" sarà identificata dalla lettera C, lo "scasso" dalla B, A rappresenterà l'allarme, E il "terremoto".

Quindi per quanto appena detto si è interessati al calcolo di P(C = *true* |B = *true*):

P(C=true|B=true)=
P(C=true|A=false)P(A=false|E=false,B=true)P(E=false)+
P(C=true|A=false)P(A=false|E=true,B=true)P(E=true)+
P(C=true|A=true)P(A=true|E=false,B=true)P(E=false)+
P(C=true|A=true)P(A=true|E=true,B=true)P(E=false)
=0,7548.

Come si può notare la complessità della (3.1) aumenta con il crescere del dominio: in particolare per un network con venti variabili aleatorie binomiali, dove una è osservata e in una seconda si ha l'evidenza, la sommatoria è composta da 2^{18} termini. Cooper (1990) ha mostrato che in un generico DAG l'inferenza è NP-hard.

Per risolvere il problema sono stati sviluppati molti approcci: un'idea è quella di ottimizzare le distribuzioni marginali fattorizzando le distribuzioni; in altri termini, inserire la sommatoria più tardi possibile. Questo nell'esempio riportato si traduce nel modo seguente:

$$P(c|b)=\sum_{a,e} P(e)P(a|e,b)P(c|a)=\sum_{e} P(e)\sum_{a} P(a|e,b)P(c|a) \tag{3.5}$$

La complessità della marginalizzazione dipende dai termini presenti nella sommatoria. Si può dimostrate che introducendo un ordine alle variabili aleatorie, che minimizza i termini della somma, la complessità risulta NP-hard. (Arnborg, 1987).

3.1 Algoritmi per l'inferenza esatta

Per risolvere i problemi appena esposti, sono stati sviluppati alcuni algoritmi per l'inferenza esatta se la rete è un **polialbero**.

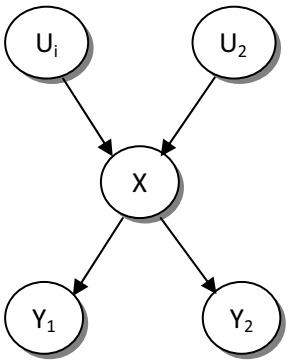


FIGURA 8: STRUTTURA DI UN POLIALBERO

Un polialbero è un grafo che ha al massimo un cammino tra due paia di nodi e si riferisce ad una singola connessione. La differenza tra un albero ed il polialbero è che il secondo può avere più di un parente.
Gli algoritmi principalmente utilizzati sono due[8]:

- ✓ Algoritmo *message passing:* si basa sul concetto di semplificazione della struttura della rete, utilizzando alberi e polialberi. Una struttura ad albero è un DAG in cui esiste solo un nodo, detto radice dell'albero che non ha genitori. Ogni altro nodo della struttura ha esattamente un parente ed è un discendente del nodo della radice. L'algoritmo venne sviluppato da Pearl (1982) utilizzando le indipendenze locali. Dato un insieme e di valori assunti dalle variabili aleatorie instanziate $E \subset X$, l'algoritmo determina $P(X_i|e)$ per tutti i valori assunti da ogni X_i della rete. Questo processo avviene inizializzando i messaggi che ogni variabile instanziata manda ai suoi nodi vicini, che a loro volta mandano ai propri vicini. L'ordine con cui vengono inizializzati i messaggi non influenza l'aggiornamento delle probabilità. L'algoritmo è stato poi generalizzato dallo stesso Pearl nel 1988, in cui la nuova struttura di riferimento è un polialbero[9].

- ✓ Algoritmo *junction tree*. Esiste un altro metodo efficiente per la propagazione dell'evidenza basata sul concetto del junction tree. In questo contesto, la rete bayesiana originale viene trasformata in un grafo indiretto, detto appunto junction tree, in grado comunque di descrivere e rappresentare la distribuzione di probabilità congiunta e quindi utilizzare la struttura trovata per fare inferenza. (Jensen, Lauritzen Spiegelhalter, 1996).

  In particolare quello che accade può essere riassunto nei seguenti punti:

  1. Implementazione della base di conoscenza
  2. Compilazione o inizializzazione del grafo
  3. Motore di inferenza con propagazione dell'inferenza

[8] Non verrà trattato interamente ogni algoritmo. Per maggiori informazioni si rimanda quindi alle trattazioni originali.
[9] La struttura ad albero è un caso particolare della struttura a polialbero.

E' possibile utilizzare l'algoritmo di propagazione dell'informazione per trovare la configurazione con massima probabilità oppure generare valori simulati dalla distribuzione di probabilità congiunta.

In letteratura sono presenti altri algoritmi per fini inferenziali :

- Cutset conditioning:

  Pearl propone questa tecnica per trattare il problema di propagazione dell'evidenza in un network con connessioni multiple. L'idea principale è trovare l'insieme minimale di nodi che verrà instanziato dal resto del network costituito da connessioni singole.

  Pearl 1997, Darwiche 1995, Suermondt and Cooper 1990.

- Gibbs sampling:

  L'algoritmo viene utilizzato specialmente nei casi cui l'informazione è incompleta, ma le distribuzioni condizionate di ogni variabile sono note. L'obbiettivo è generare un'istanza dalla distribuzione di ogni variabile condizionata allo stato degli altri nodi.

  Pearl 1987, Chavez and Cooper 1990.

- Arc reversal/node reduction:

  Egli applica una sequenza di operatori all'interno del network, che inverte i collegamenti tra i nodi utilizzando la regola di Bayes. Il procedimento continua finché il network non si riduce agli unici nodi la cui evidenza deriva dai soli nodi predecessori.

  Shachter 1990.

- Variable elimination:

  L'algoritmo elimina una variabile alla volta, sommandole. La complessità può essere misurata tramite il numero di moltiplicazioni e addizioni necessario che devono essere eseguite. L'eliminazione che produce una complessità minore è quella ottimale.

- Symbolic probabilistic inference (SPI):

  L'inferenza probabilistica viene vista come una combinazione di ottimizzazioni. Il problema è trovare la fattorizzazione ottimale data da un insieme di distribuzioni di probabilità.

  Jensen, Li and D'Ambrosio (1994).

3.2 ALGORITMI PER L'INFERENZA APPROSSIMATA

I ricercatori hanno sviluppato alcuni algoritmi alternativi per l'inferenza approssimata. Una strada possibile per operare è simulare un insieme di dati usando numeri pseudo casuali in accordo con la distribuzione di probabilità del network, ed in seguito approssimare le probabilità condizionate di interesse, usando il campione simulato. Questo metodo è noto come "stochastic simulation".
Gli algoritmi stocastici di simulazione, anche chiamati "Monte Carlo" sono gli algoritmi più conosciuti; essi generano un insieme di campioni o di network instanziati basandosi sulla probabilità del modello di partenza e poi approssimano le probabilità dei nodi tramite le frequenze campionarie.
L'insieme di questi algoritmi può essere scomposto in due parti: "importance sampling algorithms" [G.Casella, E. I. George (1992)] e "Markov Chain Monte Carlo" (MCMC) [B. Gill, G.Casella(2004)].

I metodi, che appartengono invece alla categoria "Model semplification", semplificano in primo luogo il modello fino a quando non risulta possibile usare un algoritmo per l'inferenza esatta. Di questa categoria fanno parte "Sarkar's algorithm" [V. E. Media, D.Sarkar(1992)] e "mini-buckets" [K. Kask, Dechter (1999a)].
Negli ultimi anni sono stati introdotti algoritmi appartenenti all'insieme "loopy belief propagation" che utilizzano l'algoritmo di propagazione sviluppato da Pearl in network, la cui struttura è un polialbero in cui sono presenti dei salti (loop). I ricercatori hanno dimostrato empiricamente che la loro efficienza è alta per alcuni network, tuttavia in generale i risultati sono da considerarsi più poveri e di scarsa efficienza.
L'ultima categoria, che qui viene presentata, riguarda i metodi "search based"; essi ricercano network con probabilità più alta, usando poi la struttura risultante per ottenere ragionevoli approssimazioni. Di questa categoria fanno parte: Deterministic Approximation" [E. Santos, S.E. Shimoy (1998)], "Sample-and-Accumulate" [J.R. Shimony (1996)], "Top-N" [L. Zun, W. Meng]. Druzdzel e Lin hanno empiricamente dimostrato che con questi metodi, anche con un numero esiguo di evidenze sulle variabili aleatorie, è possibile replicare buona parte della probabilità del fenomeno.

4. APPRENDIMENTO DEI NETWORK BAYESIANI

In questo capitolo verrà affrontata la fase dell'apprendimento (o learning) di una rete bayesiana. L'obbiettivo del learning è trovare la struttura che descriva al meglio il dataset iniziale: questa fase è il fulcro dell'intera analisi, infatti i risultati che ne derivano sono il punto di partenza per l'inferenza descritta nel capitolo precedente e rappresentano la descrizione delle relazioni di in/dipendenza.

Esiste un parallelismo tra la definizione di una rete bayesiana e il processo che genera la struttura; il learning infatti è costituito da due fasi che verranno sviluppate nel dettaglio: l'apprendimento dei parametri e l'apprendimento della struttura. Queste due fasi sono interconnesse tra loro, il learning della struttura spiega i dati una volta forniti i parametri adatti; l'apprendimento dei parametri è possibile se è data una struttura. Il procedimento naturale porta quindi a considerare prima l'apprendimento della struttura ed in seguito quello dei parametri.
Il processo di learning prende varie forme a seconda delle diverse situazioni in cui è svolta l'analisi: le variabili aleatorie sono completamente osservabili o meno, in altre parole se si è in presenza di dati mancanti; la struttura può essere nota o ignota; più in particolare le variabili aleatorie possono essere continue o discrete, queste ultime possono essere dicotomiche.
Si possono individuare quattro principali situazioni:

- ✓ Struttura nota e dati incompleti
- ✓ Struttura ignota e dati incompleti
- ✓ Struttura nota e dati completi
- ✓ Struttura ignota e dati completi.

E' possibile apprendere la struttura del DAG in differenti modi: analizzando i rapporti di dipendenza tra variabili aleatorie descritti da un esperto del settore oppure derivare le relazioni tramite i dati in modo automatico. Il primo approccio è spesso dispendioso e richiede molto tempo specialmente per network complicati, il secondo utilizza software che permettono di apprendere la struttura in modo automatico inserendo condizioni fornite dagli esperti. Quest'ultimo approccio è quello usato più comunemente, poiché meno dispendioso, ma efficace.
In questi ultimi anni i ricercatori hanno focalizzato la loro attenzione sullo sviluppo di tecniche che permettono di derivare direttamente dai dati la struttura della rete. L'utilizzo di tali algoritmi permette di ottenere le informazioni necessarie per apprendere le relazioni esistenti tra le variabili aleatorie appartenenti al database, inserendo le eventuali informazioni che l'analista ha disposizione.

In questa tesi verrà affrontato il caso in cui la struttura è ignota, i dati sono completi e le variabili aleatorie aleatorie sono discrete e multinomiali.

4.1 APPRENDIMENTO DEI PARAMETRI

In questa sezione occorre supporre di conoscere la struttura della rete. Si riporta di seguito un semplice esempio riferito a variabili aleatorie discrete, in cui si è interessati alla conoscenza di un singolo parametro.
Si supponga di avere 101 monete in un'urna, ciascuna con una differente probabilità di far uscire testa in un lancio. La probabilità per la prima è pari a 0.00, per la seconda è 0.01, per la terza è 0.02 e così via. Questo vuol dire che se ad esempio si lancia un numero ragionevole di volte la terza moneta, la probabilità che esca testa si aggira intorno a 0.02. Si supponga di estrarre casualmente una moneta dall'urna e lanciarla: si è quindi interessati a conoscere la probabilità che esca testa. Se si conoscesse la frequenza relativa, ossia quel numero che descrive il comportamento della variabile, la probabilità in esame sarebbe pari alla frequenza relativa associata. In altre parole, se si potesse identificare la moneta (la prima, la seconda,…etc) la stima del parametro sarebbe banale. Si introduce la variabile casuale "Lato" (L) le cui modalità sono "testa" o "croce", ed "F" la variabile composta dai 101 valori delle frequenze relative.

$$P(Lato=testa|f)=f \tag{4.1}$$

Se si assegnano uguali probabilità a tutte le frequenze relative, ossia alle monete dell'urna, è possibile rappresentare l'esempio nella rete bayesiana sottostante:

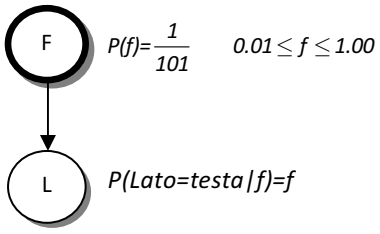


FIGURA 9: GRAFO CASUALE

La probabilità che esca testa è così definita:

$$P(Lato=testa)=\sum_{f=0}^{1} P(Lato=testa|f)P(f)=\sum_{f=0}^{1} f\left(\frac{1}{101}\right)=\left(\frac{1}{100 \times 101}\right)\sum_{f=0}^{100} f= \\ =\left(\frac{1}{100 \times 101}\right)\left(\frac{100 \times 101}{2}\right)=\frac{1}{2} \tag{4.2}$$

Non sorprende che la probabilità risultante sia 0.5, poiché le frequenze relative sono distribuite equamente da entrambi i lati rispetto 0.5.

La distribuzione più utilizzata in ambito bayesiano è quella di **Dirichlet** che può essere scritta come:

$$p(\vartheta)=\lambda\vartheta^{\alpha}(1-\vartheta)^{\beta} \tag{4.3}$$

Questa distribuzione è utile perché la distribuzione a posteriori è ottenuta a partire dalla sua forma generale e quindi è ancora una distribuzione Dirichlet[10]. In particolare questa relazione implica che la a posteriori è un semplice aggiornamento dei parametri della a priori.

$$p(\vartheta)=\lambda\vartheta^{p+\alpha}(1-\vartheta)^{f+\beta} \tag{4.4}$$

E la sua media risulterà pari a:

$$E(\vartheta)=\frac{p+\alpha}{p+f+\alpha+\beta} \tag{4.5}$$

Si supponga di avere un insieme di nodi **X** e si supponga che i nodi possano assumere solo valori discreti, si consideri inoltre un insieme di n-uple D. Se X_i è un nodo r_i è il numero dei suoi possibili risultati, c_i l'insieme dei genitori , ϑ_{ijk} è la probabilità che X_i sia nello stato k condizionatamente al fatto che l'insieme dei suoi genitori è complessivamente nello stato j.

Si supponga che l'insieme D arrivi da una struttura di un network B_s , inoltre che ϑ_{ijk} siano parametri indipendenti tra loro e siano distribuiti secondo una legge di Dirichlet; si può quindi scrivere:

$$E(\vartheta_{ijk}|D,B_s)=\frac{N_{ijk}+\alpha_{ijk}}{N_{ij}+\alpha_{ij}} \tag{4.6}$$

Dove N_{ijk} è il numero di n-uple tali che X_i è nello stato k dato che i suoi genitori sono complessivamente nello stato j; N_{ij} è il numero delle n-uple in D, tali che i genitori del nodo X_i sono complessivamente nello stato j ed inoltre indipendente dallo stato del nodo. Inoltre

$$N_{ij}=\sum_{k=1}^{r_i} N_{ijk} \tag{4.7}$$

α_{ijk} è l'esponente del parametro ϑ_{ijk} nella distribuzione di Dirichlet iniziale e $\alpha_{ij}=\sum_{K=1}^{r_i}\alpha_{ijk}$.

Devono inoltre essere soddisfatte alcune proprietà:

[10] Distribuzione coniugate.

- ✓ Markovianità: ogni nodo deve essere indipendente da tutti i suoi non discendenti condizionatamente ai suoi genitori.
- ✓ Causalità: non devono esistere variabili aleatorie latenti e devono essere distinte tra loro. Se ciò non accadesse verrebbe meno la corrispondenza biunivoca tra i nodi del grafo e le variabili aleatorie stesse.

- ✓ Accuratezza: il grafo deve essere una mappa perfetta[11] della distribuzione di probabilità considerata.

Un'ipotesi ulteriore riguarda la fattorizzazione nella probabilità secondo una certa struttura S. Si ricerca quindi l'insieme dei parametri che massimizza la log-verosimiglianza per la struttura S.

L'approccio comunemente usato è il seguente: si suppone di avere una a priori che descrive l'informazione disponibile (se l'informazione non esiste, è possibile usare una variabile casuale che segue la distribuzione di una uniforme) e scelta una distribuzione appartenente alla famiglia della a priori, detta coniugata, vengono aggiornati i parametri per determinare la distribuzione finale. In questa sede per le distribuzioni locali si utilizzeranno delle distribuzioni multinomiali. La distribuzione coniugata appartiene alla famiglia Dirichlet. Se si indica con ϑ_{ijk} la probabilità locale del nodo considerato per i genitori, la distribuzione può essere così scritta:

$$Dir(\alpha_{ij1},\alpha_{ij2},\ldots,\alpha_{ijr_i})=\Gamma(\alpha_{ij})\prod_{k=1}^{r_i}\frac{\vartheta_{ijk}^{\alpha_{ijk}-1}}{\Gamma(\alpha_{ijk})} \tag{4.8}$$

Assumendo l'indipendenza locale e globale, l'insieme dei parametri **p** del network G è:

$$Pr(\vartheta|G)=\prod_{i=1}^{n}\prod_{j=1}^{q_i}\Gamma(\alpha_{ij})\prod_{k=1}^{r_i}\frac{\vartheta_{ijk}^{\alpha_{ijk}-1}}{\Gamma(\alpha_{ijk})} \tag{4.9}$$

Condizionata al dataset la probabilità a posteriori fa parte della famiglia Dirichlet ed è coniugata alla multinomiale:

$$Pr(\vartheta|G,D)=\prod_{i=1}^{n}\prod_{j=1}^{q_i}\Gamma(\alpha_{ij}+N_{ij})\prod_{k=1}^{r_i}\frac{\vartheta_{k}^{N_{ijk}+\alpha_{ijk}-1}}{\Gamma(\alpha_{ijk}+N_{ijk})} \tag{4.10}$$

Da qui si può dimostrare che:

$$\hat{\vartheta}_{ijk}=\frac{\alpha_{ijk}+N_{ijk}}{\alpha_{ij}+N_{ij}} \tag{4.11}$$

In altri campi della statistica la stessa quantità viene calcolata tramite l'utilizzo di alcune statistiche sufficienti ed assume la seguente forma:

[11] Definizione presente nell'appendice.

$$\hat{\vartheta}_{ijk} = \frac{N_{ijk}}{N_{ij}} \tag{4.12}$$

4.2 APPRENDIMENTO DELLA STRUTTURA

Nel paragrafo precedente si è supposto di conoscere la struttura per procedere con il learning dei parametri. Nella maggior parte dei casi è necessario apprendere anche la struttura ed anche questa procedura si basa sull'approccio automatico, in altre parole tutte le informazioni necessarie per la costruzione di una struttura adeguata, sono da ricercare all'interno del dataset.
Si assume che:

- ✓ Le variabili aleatorie siano interamente osservabili[12];
- ✓ La distribuzione delle frequenze relative delle variabili aleatorie ammette una rappresentazione tramite DAG;
- ✓ La distribuzione delle frequenze relative racchiude tutte le informazioni riguardo l'insieme delle indipendenze condizionate;
- ✓ La distribuzione delle frequenze relative ammette una rappresentazione faithful del DAG.

Definizione network bayesiano multinomiale:

Le seguenti proprietà costituiscono lo schema dell'apprendimento della struttura di un network bayesiano multinomiale:

1. *n variabili aleatorie* $X_1, X_2, ..., X_n$ *con una distribuzione discreta congiunta P;*
2. *un campione di dimensione N;*
3. *una variabile casuale GP il cui range è costituito da tutti i possibili cammini contenenti le n variabili aleatorie, e per qualunque valore gp che un cammino può assumere, ed una probabilità a priori P(gp);*
4. *un insieme* $D = \left\{ X^{(1)}, X^{(2)}, ..., X^{(M)} \right\}$ *di vettori casuali n-dimensionali.*

   Per ogni valore gp di GP, D è un campione di network bayesiani multinomiali di dimensione M con parametri $(G, F^{(G)})$ *, dove* $(G, F^{(G)}, \rho \mid G)$ *è il network bayesiano con tutti i possibili cammini specificati.*

[12] Come già dichiarato, in questa tesi non si tratterà il caso di dati mancanti.

Supponendo di avere un insieme di dati riferiti ai valori del vettore in D, si può scrivere la seguente relazione:

$$P(d \mid gp) = P(d \mid G) = \prod_{i=1}^{n} \prod_{j=1}^{q_i^{(G)}} \frac{\Gamma(N_{ij}^{(G)})}{\Gamma(N_{ij}^{(G)} + M_{ij}^{(G)})} \prod_{k=1}^{r_i} \frac{\Gamma(a_{ijk}^{(G)} + s_{ijk}^{(G)})}{\Gamma(a_{ijk}^{(G)})} \tag{4.13}$$

Dove $a_{ijk}^{(G)}$, $s_{ijk}^{(G)}$ sono i valori in $G, F^{(G)}, \rho \mid G$.

Dato un network bayesiano multinomiale ed un insieme di dati, un modo per selezionare un modello è attribuire uno score ai diversi DAG, quello con punteggio più alto verrà selezionato. E' indifferente calcolare gli score nel DAG o nei cammini del DAG, infatti esiste la seguente relazione:

$$score_B(d, gp) = score_B(d, G) = P(d \mid G) \tag{4.14}$$

Si consideri un esempio in cui sono presenti due variabili aleatorie, ciascuna con due possibili modalità.

FIGURA 10: ESEMPIO POSSIBILI STRUTTURE

Si prenda in considerazione il primo cammino riportato in cui le due variabili aleatorie sono dipendenti ed il secondo in cui sono indipendenti. Si supponga inoltre di avere ottenuto i seguenti dati:

Case	X_1	X_2
1	1	1
2	1	2
3	1	1
4	2	2
5	1	1
6	2	1
7	1	1
8	2	2

TABELLA 1: ESEMPIO DATASET

E' necessario calcolare le probabilità basandosi sui dati conoscendo i due cammini possibili:

$$P(d|gp_1)=\left(\frac{\Gamma(4)}{\Gamma(4+8)}\frac{\Gamma(2+5)\Gamma(2+3)}{\Gamma(2)\Gamma(2)}\right)\left(\frac{\Gamma(2)}{\Gamma(2+5)}\frac{\Gamma(1+4)\Gamma(1+1)}{\Gamma(1)\Gamma(1)}\right)\left(\frac{\Gamma(2)}{\Gamma(2+3)}\frac{\Gamma(1+2)\Gamma(1+1)}{\Gamma(1)\Gamma(1)}\right)=$$

$$= 7{,}2150 \times 10^{-6}$$

$$P(d|gp_2)=\left(\frac{\Gamma(4)}{\Gamma(4+8)}\frac{\Gamma(2+5)\Gamma(2+3)}{\Gamma(2)\Gamma(2)}\right)\left(\frac{\Gamma(2)}{\Gamma(4+8)}\frac{\Gamma(2+5)\Gamma(2+3)}{\Gamma(2)\Gamma(2)}\right)=$$

$$= 6{,}7465 \times 10^{-6}$$

Se si assegna $P(gp_1)=P(gp_2)=0{,}5$, si può applicare il seguente teorema e poi eseguire i calcoli relativi:

$$P(gp_1|d)=\frac{P(d|gp_1)P(gp_1)}{P(d)}=\frac{7{,}2150 \times 10^{-6} \times 0{,}5}{P(d)}=\alpha(3{,}6075 \times 10^{-6})$$

$$P(X_1=1)=7/12 \qquad P(X_2=1|X_1=1)=5/7$$

$$P(X_2=1|X_1=2)=2/5$$

Le probabilità riportate sul grafo si riferiscono ai dati presenti in tabella.

$$P(gp_2|d)=\frac{P(d|gp_2)P(gp_2)}{P(d)}=\frac{6{,}7465 \times 10^{-6} \times 0{,}5}{P(d)}=\alpha(3{,}37325 \times 10^{-6})$$

Dove α è la costante di normalizzazione $\frac{1}{P(d)}$. Sostituendola con la somma delle due probabilità riportate si ottiene:

$$P(gp_1|d)=\frac{3{,}6075 \times 10^{-6}}{3{,}6075 \times 10^{-6}+3{,}37325 \times 10^{-6}}=0{,}51678$$

$$P(gp_2|d)=\frac{3{,}37325 \times 10^{-6}}{3{,}6075 \times 10^{-6}+3{,}37325 \times 10^{-6}}=0{,}48322$$

E' immediato dedurre che tra le due strutture è preferibile la prima, quella che indica dipendenze condizionate poiché è più probabile sulla base dei dati considerati.

E' possibile quindi fare inferenza tramite il DAG scelto che descrive la situazione in esame:

$$P(X_1=2|X_2=1)=\frac{P(X_2=1|X_1=2)P(X_1=2)}{P(X_2=1|X_1=1)P(X_1=1)+P(X_2=1|X_1=2)P(X_1=2)}=$$

$$=\frac{(2/5)(5/12)}{(5/7)(7/12)+(2/5)(5/12)}=0{,}28571$$

Questo esempio, pur nella sua semplicità, fornisce un'idea generale sui metodi di selezione della struttura che verranno esplicati in modo approfondito nei paragrafi che seguono.

I metodi per il learning della struttura possono essere divisi in due filoni principali:

- ✓ Search-score: come l'esempio sopra riportato, si usano funzioni score (punteggi) che permettono di confrontare l'adeguatezza tra le possibili strutture di una rete;

- ✓ Constraint-based: in questo caso si utilizzano delle misure per indagare l'esistenza di eventuali in/dipendenze condizionate tra le variabili aleatorie.

In entrambi i casi si seleziona la struttura che maggiormente si adatta all'insieme di dati fornito.

4.2.1 METODI SCORE-BASED

L'obbiettivo di questi metodi è selezionare la rete a cui è associato lo score più alto, confrontandolo secondo misure di bontà di adattamento. Quest'ultimo può essere calcolato utilizzando la log-verosimiglianza, o le sue versioni penalizzate AIC (Akaike information criterior) ed il BIC (Bayesian information criterior):

$$Score_{ML}(G,D) = \log L(D \mid G)$$
$$Score_{AIC}(G,D) = \log L(D \mid G) - d$$
$$Score_{AIC}(G,D) = \log L(D \mid G) - \frac{d}{2}\log N$$

Le ultime due versioni sono penalizzate in base al numero di parametri, in particolare vengono privilegiati i modelli più semplici a parità di precisione.

Esiste un'ulteriore strada, ossia utilizzare la distribuzione a posteriori calcolata secondo il metodo di Bayes. Questo procedimento ha lo scopo di massimizzare il seguente score:

$$Score_{Bayes}(G,D)=Pr(G|D)=\frac{Pr(D|G)Pr(G)}{Pr(D)} \tag{4.15}$$

Che può essere così approssimato:

$$Score_{Bayes}(G,D)=Pr(D|G)Pr(G) \tag{4.16}$$

Per massimizzare lo score bisogna massimizzare il numeratore poiché il denominatore dipende solo dai dati e non dalla struttura del grafo.

$$Pr(D|G)=\int Pr(D|G,p)\, Pr(p|G)\, dg \tag{4.17}$$

Ed assumendo come vera la proprietà di Markov, si può scrivere la (4.17) come prodotto delle probabilità locali.

Nel caso multinomiale la probabilità a posteriori può essere scritta in forma esplicita (operazione non sempre possibile) utilizzando la distribuzione coniugata Dirichlet, come distribuzioni a priori sullo spazio parametrico, ed assumendo l'indipendenza locale e globale dei parametri:

$$P(D \mid G) = \prod_{i=1}^{n}\left(\prod_{j=1}^{q_i}\frac{\Gamma(\alpha_{ij})}{\Gamma(\alpha_{ij}+n_{ij})}\prod_{k=1}^{r_i}\frac{\Gamma(\alpha_{ijk}+n_{ijk})}{\Gamma(\alpha_{ijk})}\right) \tag{4.18}$$

dove le lettere indicate si riferiscono alle medesime quantità della (4.13).

Gli algoritmi riportati in seguito partono da un grafo senza archi e usano un metodo di ricerca ("search") per aggiungere eventuali archi; basta poi calcolare lo score della struttura per valutare se l'introduzione degli archi è necessaria o meno nella stima del grafo. Si continua il processo finché la struttura trovata non migliora quella precedente.

Per calcolare lo score si possono utilizzare differenti metodi: scoring method [Cooper and Herskovits 1991], entropy based method [Herskovits, 1991], minimum description length method [Suzuki, 1996, Lam and Bacchus, 1994], e minimum message length method [Wallace 1996]. Talvolta per ridurre lo spazio di ricerca, molti di questi algoritmi richiedono di ordinare i nodi. Questi metodi sono NP-hard.

4.2.2 METODI CONSTRAINT-BASED

Questi metodi valutano le in/dipendenze condizionate tra variabili nel dataset.

In generale si tenta di indagare sull'esistenza della dipendenza tra due variabili aleatorie che nel grafo viene tradotta con l'introduzione di un arco tra i due nodi in esame.

La complessità computazionale richiesta per il raggiungimento dello scopo è uno dei principali problemi; per questo motivo si richiede l'introduzione di alcuni metodi euristici che rispondano ad alcuni criteri:

- ✓ minimizzare il numero di test di indipendenza condizionale, per ridurre l'incidenza di errori di prima e seconda specie;
- ✓ contenere la dimensione dei condizionamenti poiché occorre un numero sufficiente di osservazioni per ogni combinazione;
- ✓ utilizzare un test che non richieda complessità computazionali elevate.

Gli svantaggi di questi metodi riguardano la scarsa robustezza: piccole variazioni sui dati iniziali portano a rilevanti differenze nei risultati. Inoltre questi algoritmi sono esponenziali, ossia un numero elevato di variabili aleatorie implica un numero elevato di test da verificare.

4.3 ALGORITMI SCORE-BASED E CONSTRAINT-BASED

Di seguito si riporta una semplice descrizione dei più importanti algoritmi score-based[13]:

- **Sparse candidate algorithm [Friedmann(1999)]**

  L'algoritmo è stato introdotto da Friedmann (1999). L'idea di base è di ridurre lo spazio di ricerca limitando il numero dei genitori dei nodi del network bayesiano. Si introducono delle misure di mutua informazione per escludere strutture che possono essere considerate indipendenti.

  E' composto da due step: nel primo si continua a ridurre lo spazio, si parla infatti di restrict step; nel secondo, che è chiamato massimizzatore (greedy hill-climbing, metodo euristico), un algoritmo viene iterato per massimizzare lo score nello spazio ristretto precedentemente selezionato. Viene penalizzato il modello se è troppo complesso.

- **Algoritmo per la costruzione un albero [Chow-Liu, 1968]**

  Si deve scegliere una distribuzione di probabilità P come input e come output viene prodotto un albero in $O(N^2)$ passi (dove N rappresenta il numero dei nodi). L'idea generale è trovare la struttura con il miglior score; l'algoritmo termina con l'analisi di dipendenza tra coppie di variabili aleatorie. Questo algoritmo non può però essere utilizzato per la ricerca di una struttura in cui sono presenti connessioni multiple, poiché la valutazione della dipendenza viene fatta per coppie di variabili aleatorie.

- **Algoritmo Kutato [Cooper and Herskovits, 1991]**

  In questo caso come misura dello score viene utilizzata l'entropia. Il problema viene visto come approssimazione della vera probabilità di giunzione dei dati usando la rete che ha la minima informazione persa (massima entropia), inoltre è richiesta la specificazione dell'ordine con cui trattare i nodi.

- **Algoritmo BENEDICT [Acid and Campos, 1996]**

  Anche in questo caso viene richiesto un ordine dei nodi; viene utilizzato un metodo euristico per la ricerca e si utilizza un metodo basato sull'entropia. Dopo aver scelto la struttura per via euristica l'algoritmo analizza poi l'indipendenza condizionata usando il concetto della d-separazione e ne calcola la differenza con quella reale dei dati.

- **Algoritmo CB [Singh and Valtorta, 1995]**

  In questo caso l'ordine dei nodi deve essere scelto da esperti: l'algoritmo ha la capacità di determinare l'orientamento degli archi. Gli sviluppatori hanno prodotto un algoritmo ibrido che

[13] Per informazioni più dettagliate, consultare gli articoli originali.

utilizza sia l'algoritmo PC per trovare l'ordine dei nodi e poi utilizza una versione modificata del K2 per determinare la struttura. Quindi a volte il CB può anche rifiutare l'ordine dei nodi fornito dagli esperti.

Si riportano di seguito alcuni algoritmi search-based:

- **Algoritmo SGS [Spirtes, Scheines e Glymour, 1990]**

  Questo algoritmo non richiede di specificare l'ordine con cui trattare i nodi; esso orienta automaticamente gli archi della struttura usando test di indipendenza. L'algoritmo ha una complessità esponenziale.

- **Algoritmo PC [Spirtes e Glymour, 1991]**

  E' una versione aggiornata dell'SGS algoritmo; questa versione risulta più efficiente quando si costruiscono modelli con un numero di nodi non particolarmente elevato.

- **GS(greedy-search) algoritmo [Scheines e Glymour, 1988]:**

  Questa versione è valida per i casi in cui la grandezza e la quantità delle variabili aleatorie è piccola. In ogni caso l'algoritmo GS è asintoticamente più veloce di altri algoritmi anche in presenza di numerose variabili aleatorie.

  Il Markov blanket ($BL(X)$) è un punto centrale di questo algoritmo. L'algoritmo si sviluppa in due parti: una fase di crescita in cui vengono aggiunte nel Markov blanket più variabili aleatorie possibili, inserendo così anche variabili aleatorie che non dovrebbero esserci, si seleziona la variabile di cui si desidera creare il $BL(X)$, si applica un test di indipendenza per ogni variabile e si aggiungono gli archi a seconda del risultato del test. A questo punto si forma l'insieme di interesse.

  Nella seconda parte, vengono eliminate le variabili aleatorie che non dovrebbero appartenere a questo insieme e che chiaramente sono indipendenti dal resto perché d-separate dal $BL(X)$.

  Il vantaggio maggiore nell'utilizzo di tale algoritmo riguarda l'uso del $BL(X)$ che restringe la grandezza dell'insieme di condizionamento, inoltre durante la seconda fase rimuove i nodi che sono stati aggiunti erroneamente durante la computazione di Markov Blanket.

- **Algoritmo Wermuth-Lauritzen [Wermuth e Lauritzen, 1983]**

  Per ogni nodo appartenente al network si calcola un test di indipendenza con un altro nodo, se le variabili aleatorie risultano dipendenti si aggiunge un arco. In ogni caso all'aumentare dei nodi aumenta il numero di test necessari e quindi talvolta questo algoritmo non è utilizzabile.

4.3.1 L'ALGORITMO K2[14]

In questo paragrafo viene descritto nel dettaglio il funzionamento dell'algoritmo K2 che sarà utilizzato per l'analisi del dataset reale.

L'algoritmo K2 fa parte dei metodi score-based e può essere utilizzato per una qualsiasi rete bayesiana. Come input necessita di un dataset ed un insieme di nodi ordinati: il dover fornire un ordine dei nodi può essere un limite, che in alcuni casi può essere risolto con l'aiuto della conoscenza di un esperto o attraverso analisi preliminari. Il risultato è chiaramente un network che descrive la relazione tra le variabili aleatorie in esame.

Il procedimento per la ricerca delle dipendenze tra le variabili aleatorie avviene nel seguente modo: il ricercatore introduce una struttura ipotetica di dipendenza e tramite l'ausilio di un calcolatore viene individuata la probabilità della distribuzione a posteriori.

Per ogni nodo nel network bayesiano esistono delle probabilità condizionate che mettono in evidenza le relazioni tra il nodo stesso ed i suoi predecessori: ogni X_i ha un insieme di genitori che denoteremo con π_i

Si può mostrare che la probabilità congiunta di un particolare insieme di variabili aleatorie è:

$$P(X_1,X_2,.....,X_n)=\prod_{i=1}^{n} P(X_i|\pi_i) \tag{4.19}$$

Nel seguente paragrafo si utilizzerà la seguente notazione: D sarà il dataset, Z l'insieme delle variabili aleatorie,

B_{s_i} e B_{s_j} due differenti network basati però sullo stesso insieme di variabili aleatorie Z. E' chiaro che le due reti avranno una differente struttura e per poter selezionare la più adatta bisogna valutare l'adattamento di ciascuno dei due network:

$$\frac{P(B_{si}|D)}{P(B_{sj}|D)}=\frac{\frac{P(B_{si};D)}{P(D)}}{\frac{P(B_{sj};D)}{P(D)}}=\frac{P(B_{si};D)}{P(B_{sj};D)} \tag{4.20}$$

Per poter procedere con il calcolo delle quantità riportate è necessario verificare che:

Assunzione 1:

Le variabili aleatorie in esame sono discrete. Questa assunzione non preclude alcuno scopo di questa tesi, poiché tutte le variabili aleatorie che qui verranno usate sono discrete. In studi in cui sono presenti variabili

[14] Articolo originale: Gregory Cooper e Edward Herskovits del 1992, pubblicato su "Machine learning, 9" pg. 309-347.

aleatorie continue è comunque possibile utilizzare l'algoritmo dopo un'adeguata discretizzazione delle variabili aleatorie.

Assunzione 2:

Dato un modello network bayesiano, i record di un database sono indipendenti tra loro.

Assunzione 3:

Non ci sono casi in cui sono presenti valori mancanti.

Assunzione 4:

La funzione di densità $f(B_s \mid B_p)$ è una variabile casuale uniforme. Ciò implica che prima di osservare il database, è indifferente attribuire un qualunque valore probabilistico alla struttura del network bayesiano.

Sviluppo del metodo esatto:

Come già anticipato, i genitori del nodo X_i vengono identificati dal seguente simbolo: π_i. w_{ij} è il valore che ogni genitore può assumere.

Si può dimostrare che:

$$P(B_s,D)=P(B_s)\prod_{i=1}^{n}\prod_{j=1}^{q_i}\frac{(r_i-1)!}{(N_{ij}+r_j-1)!}\prod_{k=1}^{r_i}N_{jki}! \qquad (4.21)$$

Dove r_i è un possibile valore che l' i-esimo nodo x può assumere, m numero di record, q_i sono i possibili valori che può assumere il parente del nodo di X_i, N_{ijk} è pari al numero di unità statistiche della variabile X_i con valore v_{ik} e contemporaneamente il parante assume w_{ij}.

E' chiaro che la bontà di un modello si valuta anche in base al tempo necessario per poter terminare l'algoritmo. In questa parte si valuta quindi il tempo di calcolo per arrivare a definire l'equazione (4.21).

Nell'articolo menzionato gli autori hanno dimostrato che $O(m\ n\ r + tb_s)^3$ è l'ordine del tempo di esecuzione dell'algoritmo, per quanto detto prima m è pari al numero di records, n il numero di nodi, r è il numero delle modalità che le variabili aleatorie possono assumere, infine t_{bs} è il tempo necessario per calcolare la probabilità a priori di una struttura, supponendo noto il valore di N_{ijk}. Inoltre se da precedenti analisi risulta noto il numero massimo di genitori, u, che possono essere contemplati nel problema in esame, la complessità del modello si riduce a: $O(m\ n\ r\ u + t_{bs})$. Inoltre se m e u sono costanti e $O(t_{bs}) = O(u\ n\ r)$, tutta la complessità per il calcolo della (4.21) si riduce a $O(m\ n)$.

Come già dichiarato l'obbiettivo è ricercare la struttura, tra tutte quelle possibili, che massimizza la probabilità in esame; ciò si traduce nella seguente massimizzazione:

$$\max_{B_s}[P(B_s,D)]=c\prod_{i=1}^{n}\max_{\pi_i}\left[\prod_{j=1}^{q_i}\frac{(r_i-1)!}{(N_{ij}+r_j-1)!}\prod_{k=1}^{r_i}N_{jki}!\right] \quad (4.22)$$

Metodo euristico

Si nota che la complessità del calcolo per la ricerca di un network bayesiano ha una complessità flessibile che si riduce facendo alcune assunzioni:

- ✓ L'esistenza di un ordine per i nodi del network;
- ✓ L'esistenza di un numero massimo di genitori per ogni nodo;
- ✓ $P(\pi_i \rightarrow x_i)$ e $P(\pi_j \rightarrow x_j)$ sono marginalmente indipendenti per $i \neq j$.

La seconda assunzione non può sempre essere verificata e per questo motivo è stato costruito un algoritmo euristico con una complessità polinomiale, che non richieda il numero massimo di genitori per ogni nodo. Viene utilizzato in questa fase un algoritmo greedy-search: inizialmente i nodi di un grafo non sono connessi tra loro, in seguito si selezionano le possibili connessione e si sceglie quella che massimizza la funzione score.

Anche in questa seconda versione sono necessarie delle restrizioni e delle assunzioni, in particolare viene richiesto l'ordine dei nodi presentii nel network, a priori qualunque struttura è considerata equivalente. L'obbiettivo, anche per questa seconda versione, è massimizzare la probabilità (4.21).

Viene riportato lo pseudo-codice dell'algoritmo:

```
1.  procedure K2
2.  {input: un insieme di n nodi, in un dato ordine, con un massimo numero u di genitori per ogni
3.           Nodo ed un database che contiene i possibili m casi.}
4.  {output: per ogni nodo viene fornito un insieme di genitori per ogni nodo}
5.  for  i:=1 to  n do
6.     π_i:=0;
7.     P_old:=g(i,π_i);
8.     OKToProceed := true
9.     While OKToProceed and | π_i |< u  do
10.        sia z un nodo in Pred(X_i)-π_i  che massimizza  g(i,π_i ,∪ {z})
11.        P_new:=g(i,π_i ,∪ {z}) ;
12.        If P_new > P_old then
13.           P_new:=P_old ;
14.              π_i := π_i ∪{z};
15.        Else OKToProceed := false;
16.    end{while};
17.    Write('Node:', X_i , 'Genitori di questo  nodo:', π_i )
18. end{for};
19. end{K2
20. };
```

Il primo passo pone l'attenzione su un nodo che non ha genitori e prosegue poi aggiungendone uno alla volta modificando così la probabilità dell'intero network. L'operazione termina quando la modifica della struttura non è più significativa. Per poter calcolare la probabilità si usa la seguente funzione (presente anche nelle pseudo-codice riportato):

$$g(i,\pi_i)=\prod_{j=1}^{q_i}\frac{(r_i-1)!}{(N_{ij}+r_j-1)!}\prod_{k=1}^{r_i}N_{ijk}! \tag{4.23}$$

Chiaramente l'algoritmo K2 è solo uno dei possibili metodi utilizzabili per massimizzare l'equazione (4.21).
Ci si deve ora focalizzare sulle probabilità derivate da un database e un network.

La probabilità condizionata attesa di un network:

Sia $\vartheta_{ijk}=P(X_i=v_{ik}|\pi_i=w_{ij})$ ossia la probabilità che X_i assuma valore v_{ik} sapendo che i genitori assumono un determinato valore w_{ij}. ϑ_{ijk} è chiamata probabilità condizionata del network. Sia A_4 l'assunzione che richiede che $f(B_p|B_s)$ sia una uniforme. Si consideri ora

$$E[\vartheta_{ijk}|D,B_s,A_4]=\frac{N_{ijk}+1}{N_{ij}+r_i} \tag{4.24}$$

chiamata stima bayesiana di ϑ_{ijk}.

$$Var[\vartheta_{ijk}|D,B_s,A_4]=\frac{(N_{ijk}+1)(N_{ij}+r_i-N_{ijk}-1)}{(N_{ij}+r_i)^2(N_{ij}+r_i+1)} \tag{4.25}$$

L'algoritmo è stato infine testato sulla rete ALARM: questo network è stato costruito per una ricerca iniziale, come prototipo per osservare problemi eventuali in una sala d'anestesia, da Beinlich (1989).
I due autori hanno generato i casi da questo network usando la tecnica di Monte Carlo (procedura priva di distorsione, ossia la probabilità che un particolare caso sia generato è uguale alla probabilità che il caso esista sulla base del problema descritto dal network). Ogni caso è costituito dai trentasette valori, uno per ciascuna variabile. L'insieme dei 10000 record viene usato come database di input per valutare l'algoritmo K2; si necessita di un ordine dei trentasette nodi che parzialmente ripete quello fornito dal network originale.
Da questi record simulati l'algoritmo K2 ha generato un network praticamente identico a quello originale, fatta eccezione per un arco omesso e un arco extra. Un'analisi più accurata mostra che l'arco tralasciato non è fortemente supportato dai dati. Il tempo per l'esecuzione dell'algoritmo è di 17 minuti circa. Nella tabella sottostante sono riportati i risultati dell'esecuzione dell'algoritmo a partire da database di dimensioni differente, ossia prendendo l'insieme di dati di dimensione inferiore a partire dai 10000 generati.

Case	Archi mancanti	Archi extra	Tempo esecuzione (sec)
100	5	33	19
200	4	19	29
500	2	7	55
1000	1	5	108
2000	1	3	204
3000	1	1	297
10000	1	1	998

TABELLA 2: EFFICIENZA ALGORITMO K2

Prendendo un insieme molto più piccolo di quello originale, in particolare pari a 3000 record, il risultato è il medesimo ed il tempo di esecuzione si riduce considerevolmente. L'algoritmo K2 è sensibile all'ordine in cui i nodi vengono inseriti e quindi analizzati.

4.3.2 L'ALGORITMO BNPC

BNPC è il secondo algoritmo che si prende in esame e appartiene ai metodi di apprendimento constraint-based. Esso è stato sviluppato da Jie Cheng[15], David Bell e Weiru Liu[16].

Esistono due versioni dell'algoritmo: una in cui è necessario introdurre l'ordine dei nodi appartenenti al dataset, l'altra che non richiede invece alcuna specificazione. In questa tesi si analizza il secondo approccio, ossia quello in cui non è necessario conoscere l'ordine dei nodi.

L'algoritmo si sviluppa in tre fasi:

1. Costruzione di una bozza: viene calcolata per ogni coppia di nodi la mutua informazione, come descritto precedentemente e si crea, sulla base delle informazioni dedotte, una bozza (un semplice grafo connesso, ma senza la direzione degli archi).

   Per quantificare l'informazione presente si utilizza la **mutua informazione**: in un network bayesiano, se due nodi sono dipendenti, conoscendo il valore di un nodo è possibile ottenere delle informazioni sul probabile valore assunto dall'altro nodo. Questa informazione è quantificabile,

[15] Dept. of Computing Science University of Alberta.
[16] Faculty of Informatics, University of Ulster.

appunto, tramite la mutua informazione. La quantità introdotta, oltre ad indicare se i nodi sono dipendenti, è in grado di quantificare l'eventuale dipendenza. La mutua informazione tra due variabili aleatorie è così definito:

$$I(X_i, X_j) = \sum_{x_i, x_j} P(x_i, x_j) \log \frac{P(x_i, x_j)}{P(x_i)P(x_j)} \tag{4.25}$$

E la mutua informazione condizionata è definita:

$$I(X_i, X_j | c) = \sum_{x_i, x_j} P(x_i, x_j, c) \log \frac{P(x_i, x_j | c)}{P(x_i | c)P(x_j | c)} \tag{4.26}$$

Dove *c* sta per insieme di nodi. Le due equazioni riportate sono stimate usando le frequenze relative calcolate sulla base dei dati presenti nel dataset.

Non inserire l'ordine dei nodi mette in evidenza due problemi: determinare quali nodi sono in/dipendenti e rilevare la direzione degli archi.

Per il primo problema i programmatori hanno sviluppato un metodo che permette di identificare quali nodi sono "collider[17]"e quali non lo sono, confrontato in termini quantitativi i test di indipendenza condizionata. Da notare che questi test sono gli stessi che vengono utilizzati per evidenziare relazioni di in/dipendenza, ma sono usati in maniera differente. Il risultato di questi test viene utilizzato anche per definire la direzione dell'arco. Questo procedimento permette di analizzare il dataset senza aumentare esponenzialmente i tempi di analisi.

2. Nella seconda fase vengono aggiunti gli archi al grafo quando una coppia di nodi risulta, dai test di indipendenza condizionata, essere dipendente. Quindi il risultato della seconda fase è un grafo in cui sono presenti tutti gli archi tra i nodi che sono risultati dipendenti, sotto l'ipotesi che il DAG in questione sia faithful. Per quest'ultima ipotesi basta far riferimento al teorema di Meek, riportato al paragrafo 2.2.
3. La terza fase utilizza il grafo che contiene esattamente tutti gli archi derivanti dalla fase precedente: vengono analizzate tutte le relazioni di dipendenza e quelle di indipendenza mediante i test IC[18] ed in base alle informazioni dedotte dal risultato dei test, viene modificato il grafo. L'ultima operazione necessaria consiste nell'aggiungere la direzione degli archi.

[17] Definizione in appendice.

[18] Test di indipendenza condizionale.

Si riporta lo pseudo-codice dell'algoritmo:

---

1. Costruzione di un grafo vuoto $G=(V,A)$ Inizializzazione di una lista L vuota.
2. Calcolo per ogni coppia di nodi $I(X_i, X_j)$ ed inserimento dell'ordine dei nodi nella lista
   Creazione di un punto p partendo dalla prima coppia di nodi nella lista
3. Aggiunta un arco in G in corrispondenza della prima coppia di nodi e spostamento del punto alla coppia successiva.
4. Procedere fino al termine della lista L.
5. Riportare p all'inizio della lista.
6. Richiamare la procedura try_to_separate_A (current graph,node1, node2).
7. *Se* i node1, node2 sono dipendenti *allora* aggiungere un arco e portare p alla coppia successiva.
8. Ritornare al punto 6 fino al termine della lista L.
9. Per ogni arco, se esistono altri cammini tra i due nodi, rimuoverli temporaneamente l'arco e richiama la procedura try_to_separate_A (current graph,node1, node2). *Se* i due nodi non possono essere separati, *allora* aggiungere l'arco, *in caso contrario* rimuoverlo in modo permanente.
10. Per ogni arco, se esistono altri cammini tra i due nodi, rimuoverli temporaneamente l'arco e richiama la procedura try_to_separate_B (current graph,node1, node2). *Se* i due nodi non possono essere separati, *allora* aggiungere l'arco, *in caso contrario* rimuoverlo in modo permanente.
11. Richiamare la procedura orient_edges (current graph)
12. Fine BNPC.

Si riporta ora lo pseudo codice relativo alle procedure richiamate.

try_to_separate_A (current graph, node1, node2):

---

1. Ricercare i vicini dei nodo1, nodo2 che sono adiacenti al cammino tra i due nodi. Riportare questi in due insiemi N1 e N2 rispettivamente.
2. Rimuovere da N1 e N2 i figli del nodo1 e nodo2 rispettivamente.
3. Se la cardinalità di N1 è maggiore di quella di N2 scambiare N1 con N2.
4. Usare N1 come insieme di nodi condizionati C.
5. Utilizzare $I(X_i, X_j|c)$ =v : *se* $v \leq \varepsilon$ *ritorna* "separati"
6. *Se* C contiene solo un nodo, *allora* vai al passo 8; *in caso contrario* $C_i = C|\{il\ nodo\ i_{th}\ di\ C\}$, $v_i = I(nodo1, nodo2|C)$. Ricercare $v_m = min(v_i)$
7. *Se* $v_m \leq \varepsilon$ ritorna "separati"; *in caso contrario* vai al passo 8 e $v_m = v, C_m = C$, vai allo step 6.
8. *Se* N2 non è stato utilizzato *allora* usa N2 come C e vai allo step 5, *in caso contrario ritorna* "fallito".

---

try_to_separate_B (current graph, node1, node2):

1. Ricercare i vicini dei nodo1, nodo2 che sono adiacenti al cammino tra i due nodi. Riportare questi in due insiemi N1 e N2 rispettivamente.
2. Ricercare i vicini dei nodo1 che sono adiacenti al cammino tra i due nodi. Riportare questi nodi in N1'.
3. Ricercare i vicini dei nodo2 che sono adiacenti al cammino tra i due nodi. Riportare questi nodi in N2'.
4. *Se* la cardinalità di N1+N1' è minore di quella di N2+N2' *allora* poni C= N1+N1', *altrimenti* C= N2+N2'.
5. Utilizzare $I(X_i, X_j | c)$ =v : *se* $v \leq \varepsilon$ *ritorna* "separati". Se C contiene solo un nodo, *allora* ritorna "fallito".
6. Sia $C'=C$. $\forall\, i \in [1, C]$ sia $C_i = C | \{il\ nodo\ i_{th}\ di\ C\}$, $v_i = I(nodo1, nodo2 | C)$. *Se* $v_i \leq \varepsilon$ *ritorna* "separato" *altrimenti* se $v_i \leq v+e$ *allora* $C' = C' | \{il\ nodo\ i_{th}\ di\ C\}$
7. *Se* la cardinalità di C' è minore della cardinalità di C *allora* sia $C=C'$; *altrimenti* ritorna "Fallito".

orient_edges (current graph):

1. $\forall$ coppia di nodi s1 e s2 che non sono connessi direttamente e è presente alla fine un nodo che è vicino sia di s1 che di s2, ricercare i vicini di s1, s2 che sono adiacenti al cammino tra i due nodi. Riportare questi in due insiemi N1 e N2 rispettivamente.
2. Ricercare i vicini dei nodi in N1 che sono adiacenti al cammino tra s1 e s2 ed inserirli in N1'.
3. Ricercare i vicini dei nodi in N2 che sono adiacenti al cammino tra s1 e s2 ed inserirli in N2'.
8. *Se* la cardinalità di N1+N1' è minore di quella di N2+N2' *allora* poni C= N1+N1', *altrimenti* C= N2+N2'.
4. Utilizzare $I(s1,s2|c)$ =v : *se* $v \leq \varepsilon$ *allora* vai al passo 8; altrimenti *se* C contiene solo un nodo, siano s1 e s2 genitori del nodo in C, vai al passo 8.
9. Sia $C'=C$. $\forall\, i \in [1, C]$ sia $C_i = C|\{il\ nodo\ i_{th}\ di\ C\}$, $v_i = I(s1,s2|C)$. *Se* $v_i \leq v+e$ *allora* $C'=C'|\{il\ nodo\ i_{th}\ di\ C\}$, sia s1 e s2 parenti del nodo i di C se il nodo i è vicino sia di s1 che di s2. $v \leq \varepsilon$ *allora* vai al passo 8.
5. *Se* la cardinalità di C' è minore della cardinalità di C *allora* sia $C=C'$ allora sia C=C', se la cardinalità di C, vai allo step 5.
6. Vai allo step 1 e ripeti finchè tutte le coppie di nodi sono state esaminate.
7. Per ogni tre nodi a, b, c, se a è genitore di b, b e c sono adiacenti e a e c non sono adiacenti e l'arco (b,c) non è orientato, sia b un genitore di c.
8. Per ogni arco tra a e b che non è orintato, se c'è un cammino diretto daa a b sia a parente di b.
9. Vai allo step 7 e ripeti finchè tutti gli archi non sono orientati

5 ANALISI DI UN INSIEME DI DATI REALI

La fase che in questa tesi si vuole analizzare è l'apprendimento di un rete bayesiana dato un insieme di dati reali. In particolare si intende mettere a confronto la struttura di un network bayesiano che deriva da un algoritmo score-based (algoritmo K2) e da uno constraint-based (algoritmo BNPC).

Nel primo caso si ricercherà la rete bayesiana a cui è associato lo score più elevato, nel secondo si valuteranno le relazioni di in/dipendenza tra le variabili aleatorie in esame mediante l'utilizzo di test di indipendenza.

Probabilmente i risultati più accurati deriveranno dal primo approccio rispetto a quello dei metodi constraint-based [Cooper and Herskovits, 1993; Acid and De Campos, 2003]. I metodi basati sui test di indipendenza potrebbero omettere archi tra nodi correlati oppure non individuare il corretto orientamento degli archi all'interno del network [Cheng et al., 2002].

L'approccio score-based permette di rilevare la struttura più probabile in cui la direzione degli archi è assegnata dall'algoritmo stesso dato l'insieme di dati [Cooper and Herskovits, 1992; Heckerman, 1998; Friedman and Goldszmidt, 1996]. Inoltre utilizzando lo score di Bayes[19] non si incorre nel problema della sovrastima dei dati [Hartemink et al., 2001].

5.1 DATASET

OMS è l'anagramma dell'Organizzazione Mondiale della Sanità, agenzia dell'ONU istituita il 7 Aprile 1948: tale organizzazione monitora il livello mondiale di salute al fine di migliorarlo e renderlo più alto possibile.

La salute è così definita: *stato di completo benessere fisico, psichico e sociale, e non di semplice assenza di malattia*.

L'indagine World Health Survey, di cui i dati fanno parte, è stata intrapresa dall'OMS nel 2002 con lo scopo di fornire dati empirici sullo stato di salute nazionale per monitorare i sistemi sanitari dei diversi paesi ed esaminare come la popolazione percepisce il proprio stato di salute.

Lo studio è stato effettuato in 74 paesi ed include donne e uomini con un'età superiore ai 18 anni che vivono in famiglia; vengono quindi esclusi dal campionamento coloro che non appartengono a nuclei familiari.

Il campionamento è di tipo probabilistico e stratificato a più stadi. In particolare gli strati sono dodici: l'area geografica, lo stato socio-economico, la presenza di centri di assistenza.

[19] Di cui si è discusso nel capitolo precedente.

I 74 paesi sono stati raggruppati in sei aree:

- AFRO (African Region Office): paesi appartenenti all'Africa,
- AMRO (American Region Office): paesi appartenenti all'America,
- EURO (European Region Office): paesi appartenenti all'Europa,
- EMRO (Eastern Mediterranean Region Office): paesi appartenenti al Mediterraneo Orientale,
- SEARO (South-East Asian Region Office): paesi appartenenti al Sud Est Asiatico,
- WPRO (Western Pacific Region Office): paesi appartenenti al Pacifico Occidentale.

Nella seguente tabella sono riportati tutti i paesi con la relativa suddivisione:

AREE GEOGRAFICHE						
AFRO	AMRO	EURO		EMRO	SEARO	WPRO
Burkina Faso	Brazil	Austria	Bosnia	Marocco	Bangladesh	Australia
Ciad	Chile	Belgium	Croatia	Pakistan	India	Japan
Comoros	Rep.	Denmark	Rep Czech	Tunisi	Myanmar	China
Congo	Dominican	Finland	Estonia	Emirati uniti	Nepal	Lao
Costa	Ecuador	France	Georgia		Sri Lanka	Malaysia
d'avorio	Guatemala	Germany	Hungary			Philippines
Ethiopia	Mexico	Ireland	Israel			Vietnam
Ghana	Paraguay	Italy	Kazakistan			
Kenya	Uruguay	Luxembourg	Lettonia			
Malawi		Netherlands	Norway			
Mali		Portugal	Romania			
Mauritania		Sveden	Russian			
Mauritius		United	Fed			
Namibia		Kingdom	Slovakia			
Senegal			Slovenia			
South Africa			Spain			
Swaziland			Turkey			
Zambia			Ukraine			
Zimbabwe			Yugoslavia			

TABELLA 3: SUDDIVISIONE PAESI IN AREE GEOGRAFICHE

Quattro di questi paesi (Cile, Romania, Yugoslavia, Giappone) sono stati esclusi dallo studio per l'indisponibilità dei dati.

Ogni unità statistica è stata sottoposta a due questionari [20], uno individuale e uno familiare. Per questa analisi si è preso in considerazione esclusivamente il questionario individuale suddiviso in nove macro-aree, ognuna delle quali indaga un aspetto diverso della vita quotidiana delle unità statistiche intervistate.

Sono state prese in esame tre macro-aree[21]:

[20] Consultabili liberamente sul sito dell'OMS.

- Variabili socio-demografiche (Respondent's Socio Demographic Characteristics):
    1. Paese di appartenenza (Country)
    2. Sesso (Q1001)
    3. Età (Q1002)
    4. Stato matrimoniale (Q1008)
    5. Livello di educazione (Q1009)
    6. Lavoro corrente (Q1012)
    7. *Ragione eventuale non lavoro (Q1014)*[22]
    8. Peso attribuito ad ogni unità statistica in base al disegno di campionamento (peso)
- Variabili sullo stato di salute (Health State Descriptions):
    9. Percezione stato di salute (Q2000)
    10. Difficoltà nelle svolgimento di attività lavorative (Q2001)
    11. Difficoltà nei movimenti (Q2010)
    12. Difficoltà ad eseguire attività vigorose (Q2011)
    13. Difficoltà nel vestirsi/lavarsi (Q2020)
    14. Difficoltà nel eseguire cure personali (Q2020)
    15. Quantità dolori (Q2030)
    16. Presenza dolori fisici (Q2031)
    17. Difficoltà di concentrazione (Q2050)
    18. Difficoltà ad imparare concetti nuovi (Q2051)
    19. Difficoltà nelle relazioni interpersonali (Q2060)
    20. Difficoltà nella gestione di conflitti e tensioni (Q2061)
    21. Indossa occhiali o lenti a contatto (Q2070)
    22. Difficoltà nel vedere e riconoscere persone incontrate per strada (Q2071)
    23. Difficoltà riconoscere oggetti (Q2072)
    24. Problemi di insonnia (Q2080)
    25. Percezione di sensazione di stanchezza (Q2081)
    26. Sensazioni di tristezza o depressione (Q2090)
    27. Sensazioni di ansia o preoccupazione (Q2091)
- Variabili riguardo a fattori di rischio (Risk Factors)
    28. Quanti frutti al giorno consumati (Q4020)
    29. Quantità di verdura consumata al giorno (Q4030)

[21] Le modalità delle variabili sono descritte nel dettaglio nell'appendice.

[22] Le variabili riportate in corsivo sono state eliminate, in seguito è presente la spiegazione dettagliata della decisione.

30. Quanti giorni di attività fisica vigorosa/sportiva (Q4031)
31. Quanti giorni di attività fisica moderata (Q4033)
32. Quanti giorni di attività fisica leggera/camminate (Q4036)
33. Tipi di pavimento nell'abitazione (Q4040)
34. Tipo di pareti nell'abitazione (Q4041)
35. Tipo di sorgente d'acqua potabile (Q4042)
36. Distanza dalla sorgente (Q4043)
37. Possibilità di avere 20 litri d'acqua potabile (Q4044)
38. Tipo di sanitari presenti nell'abitazione (Q4045)
39. Distanza di sanitari dall'abitazione (Q4046)
40. Che tipo di risorsa viene utilizzata per cucinare (Q4047)
41. Tipo di fornelli (Q4048)
42. Dove si cucina abitualmente (Q4049)
43. Possibilità di riscaldare l'abitazione (Q4050)
44. *Tipo di riscaldamento utilizzato, se presente (Q4051)*
45. *Tipo di fornelli utilizzati, se presenti (Q4052)*.

Delle quarantacinque variabili iniziali, quelle riportate in corsivo sono state eliminate a causa di una presenza massiccia di dati mancanti. In particolare per la variabile Q1014 i dati mancanti ammontano a 147947, per la Q4051 le osservazioni mancanti erano 151778 ed infine 215684 unità statistiche non hanno fornito alcuna risposta alla variabile Q4052.

Delle quarantadue variabili rimanenti, solamente due sono state modificate: la variabile età, che originariamente era riportata in anni, è stata suddivisa in classi; entrambi i software utilizzati richiedevano l'operazione di ricodifica.

La variabile "peso" era originariamente una variabile continua costruita come:

$$peso = \frac{1}{probabilità\ di\ entrare\ nel\ campione}$$

La probabilità di entrare nel campione corrisponde al rapporto tra la dimensione campionaria e la dimensione della popolazione in esame. La variabile continua è stata poi tramutata in una mutabile descritta da cinque modalità: molta alta, alta, media, bassa, molto bassa.

L'attenzione ricade sulla variabile Q2000 che descrive lo stato di percezione della salute: obbiettivo risulta quindi individuare la struttura della rete derivante dai due algoritmi utilizzati; si vuole cioè mettere in evidenza quali variabili influenzano la percezione dello stato di salute nei due casi.

Si è deciso di suddividere il campione in due sottocampioni: il primo per mettere in relazione la percezione di salute e le variabili relative allo stato di salute, il secondo invece ricerca la struttura composta dalla variabile d'interesse e i fattori di rischi a cui sono sottoposti i soggetti. In entrambi i casi le variabili anagrafiche sono state prese in esame.

Il dataset iniziale era composto da 265890; dopo aver analizzato le unità statistiche è emerso che se un soggetto non aveva dato risposta ad una domanda, in realtà ad esso corrispondevano molti dati mancanti, ossia ogni soggetto non aveva risposto a più domande. Le unità statistiche per cui non è presente alcun dato mancante sono 193057. Il numero di informazioni complete sono considerate sufficienti per proseguire l'analisi, quindi le osservazioni incomplete sono state rimosse dall'analisi.

La tabella riportata di seguito riporta il numero di variabili, l'ammontare delle unità statistiche per ogni area geografica e per ogni sottocampione costruito:

Aree Geografiche	Stato di salute		Fattori di rischio	
	Variabili	Unità statistiche	Variabili	Unità statistiche
AFRO	25	39702	23	16383
AMRO	25	44652	23	2668
EURO	25	30125	23	3471
EMRO	25	11345	23	2616
SEARO	25	31315	23	21708
WPRO	25	16383	23	8524

TABELLA 4: SUDDIVISIONE DATASET REALE

Per ogni area geografica sono stati appresi quattro network bayesiani:

1. Variabili riguardo lo stato di salute mediante l'utilizzo dell'algoritmo K2;
2. Variabili riguardo i fattori di rischio mediante l'utilizzo dell'algoritmo K2;
3. Variabili riguardo lo stato di salute mediante l'utilizzo dell'algoritmo BNPC;
4. Variabili riguardo i fattori di rischio mediante l'utilizzo dell'algoritmo BNPC.

5.2 WEKA E BNPC

L'analisi è stata effettuata utilizzando due software differenti: Weka per poter apprendere la struttura dei network bayesiani utilizzando un metodo score-based e il relativo algoritmo K2; l'algoritmo BNPC è invece implementato nel software BNPC, ed è stato quindi possibile applicare il learning per la struttura delle reti con un approccio constraint-based.

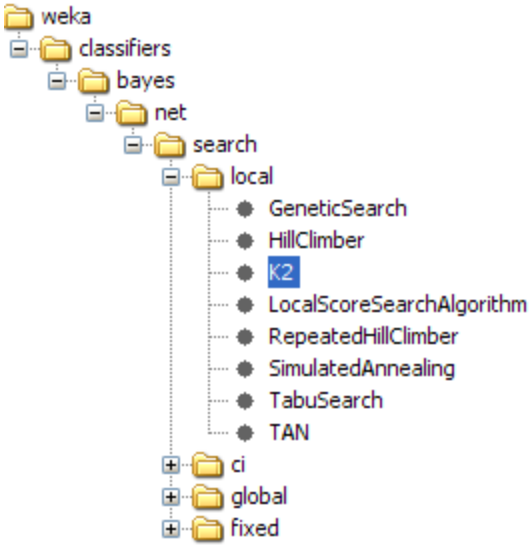
Weka è stato sviluppato dall'università di Waikato in Nuova Zelanda ed è open source. WEKA sta per "**W**aikato **E**nvironment for **K**nowledge **A**nalysis". La sigla corrisponde al nome di un animale simile al kiwi, presente solo nelle isole della Nuova Zelanda ed è rappresentato nell'immagine riportata.

WEKA è scritto interamente in Java ed è utile per applicare metodi di apprendimento automatici a partire da un insieme di dati, anche di grandi dimensioni, e di determinare quali relazioni di in/dipendenza sono rilevanti. Le informazioni che ne derivano possono poi essere utilizzate per fare previsioni o per aiutare gli esperti a prendere decisioni in ambito di incertezza.

Le variabili appartenenti allo studio devono essere discrete ed interamente note.

L'algoritmo di apprendimento della rete bayesiana si svolge in due stadi: in primo luogo viene ricercata la struttura della rete e viene poi appresa la tabella di probabilità.

L'apprendimento della struttura può essere considerato come un problema di ottimizzazione, dove la probabilità della struttura considerata sulla base dei dati, deve essere massimizzata. Una delle proprietà dei network bayesiani riguarda la scomposizione della probabilità congiunta in probabilità locali rendendo il calcolo meno oneroso: l'analisi della probabilità di un intero network, si tramuta nell'analisi di probabilità di ogni singolo nodo. Per questo motivo si parla di punteggio locale.



E' possibile sfruttare alcune opzioni presenti nel programma Weka:

- *Init As Naive Bayes*: di default è settato "true; ciò implica che il network iniziale è completo, ogni variabile è collegata con le altre. Se al contrario si sceglie "false", verrà utilizzata una struttura vuota, senza archi. Come esplicato nell'articolo riportato, l'analisi parte da un network vuoto senza archi definiti.
- *Markov Blanket Classifier*: "false" è la scelta di default. In caso contrario ogni nodo presente alle estremità della struttura viene inserito nel Markov Blanket aggiungendo un arco.
- *Score Type*: determina lo score da utilizzare per valutare lo score locale(bayes, MDL, entropy, AIC, BDeu).[23] Chiaramente viene scelto lo score di bayes.

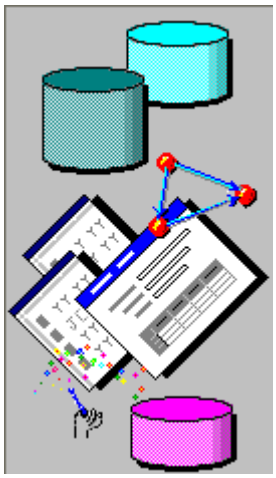
[23] Nel documento "*BouckaertBayesian Network Classifiers in Weka for Version 3-5-6",* sono presenti in dettaglio le formule relative il calcolo del score.

- *Max Nr Of Parents*: è possibile inserire il numero massimo di genitori possibili per ogni nodo. Questa opzione permette di ridurre i tempi di elaborazione dell'analisi.[24]
- *Random Order*: gli archi che connettono le variabili vengono aggiunti seguendo un ordine fisso delle stesse. Nelle opzioni è possibile richiedere che l'ordine sia casuale e che venga scelto all'inizio dell'esecuzione dell'algoritmo. Di default l'ordine randomico è settato sul "falso", quindi viene usato quello descritto nel dataset. Se l'opzione descritta in precedenza (initAsNaiveBayes) è attivata, il network iniziale terrà la variabile "classe" come la prima dell'elenco. E' quindi importante valutare inizialmente il dataset e poi stabilire quale comportamento adottare. In questo caso si è deciso di adottare un ordine randomico.

Per ogni analisi si è deciso di applicare la "cross validation[25]" con dieci incroci.

Da ogni analisi di questo tipo il software produce una struttura a cui è associato lo score più alto, ossia partendo dai dati, viene selezionata la struttura delle reti bayesiane più probabile.

Per poter apprendere la struttura seguendo un approccio constraint-based è stato utilizzato il software BNPC. Esso è stato sviluppato da Jie Cheng[26], David Bell, Weiru Liu[27]. La versione utilizzata è la 2.2 del 1997.



Il sistema del software si basa su tre fasi per poter apprendere la struttura di un network bayesiano. Così come il software presentato in precedenza, BNPC oltre ad apprendere la struttura del network apprende anche i parametri relativi. Questo aspetto però non è preso in relazione in questa sede, poiché l'attenzione viene posta esclusivamente su gli algoritmi relativi al learning della struttura.

Il software offre vantaggi e per questo è stato scelto :

- Possibilità di inserire informazione riguardo al dominio, ossia riguardo le variabili. E' infatti possibile inserire un ordinamento completo dei nodi oppure un ordinamento parziale; definire le direzioni di eventuali relazioni di cause ed effetto; specificare tra quali variabili non è possibile inserire archi; determinare quali nodi sono radici[28] e quali nodi terminali[29] del network bayesiano. Introducendo queste informazioni, se note, il tempo di apprendimento della rete diminuisce notevolmente.

[24] E' stato dimostrato nel paragrafo riguardo l'algoritmo K2.
[25] Definizione in appendice.
[26] Dept. of Computing Science University of Alberta.
[27] Faculty of Informatics, University of Ulster.
[28] Nodo senza genitori.
[29] Nodo senza figli.

- E’ possibile analizzare insiemi di dati anche molto grandi. Questo particolare aspetto è importante in questa sede, poiché come si è visto gli insiemi di dati che si andranno ad analizzare sono composti da molte unità statistiche.
- Ogni analisi genera un log file che può essere riutilizzato in un secondo momento per lo stesso dataset, rendendo così lo sviluppo dell’algoritmo molto più immediato. Dopo l’elaborazione è possibile salvare la struttura della rete e i parametri appresi in diversi formati utili per poter analizzare i risultati mediante altri software che analizzano la struttura, ad esempio Hugin (hea), Netica(nsc), Weka (bif).
- Nel pacchetto scaricabile gratuitamente è inclusa anche la possibilità di modificare il dataset grazie alla sezione “Data PreProcessor”.

E’ possibile utilizzare due differenti algoritmi a seconda che l’ordine dei nodi venga fornito o meno. Entrambi gli algoritmi sono sviluppati su tre fasi: la costruzione di una bozza, l’aggiunta degli archi mancanti e la verifica delle dipendenze mediante test di in/dipendenza condizionata. La struttura degli algoritmi è la medesima, quello che cambia è la scelta dell’algoritmo utilizzato. Per questa analisi si è deciso di non fornire alcun ordine delle variabili.

Per poter utilizzare il programma è necessario che alcune condizione siano verificate: le variabili devono essere discrete e complete; le osservazioni devono essere tra loro indipendenti data la struttura probabilistica dei dati; il dataset deve essere sufficientemente grande da rendere efficienti i test di in/dipendenza condizionata. In questo contesto tutte le condizioni sono verificate, le variabili in esame sono complete, discrete (o comunque presentano un numero finito di modalità).

L’algoritmo richiede generalmente un tempo di esecuzione dell’ordine $O(N^4)$.

5.3 Analisi dei risultati

In questo paragrafo sono riportati i risultati delle analisi.

Per ogni area geografica sono riportati quattro network bayesiani: i primi mettono in relazione “la percezione dello stato di salute” con le variabili socio-demografiche e con quelle relative allo stato di salute; di questi, il primo è riferito alla struttura appresa mediante il software Weka utilizzando il K2, ed il secondo utilizzando l’algoritmo BNPC. E’ poi analizzato il dataset in cui sono presenti i fattori di rischio: anche in questo caso il primo network riportato si riferisce alla struttura appresa dal software Weka ed il secondo riferito al software BNPC.

Per ogni area geografica sarà messa in evidenza la variabile di interesse (percezione individuale dello stato di salute), il relativo markov blanket a cui sono stati aggiunti anche i nodi che rappresentano i genitori dei genitori della variabile d'interesse.

Lo scopo è duplice: valutare, nelle diverse aree geografiche, quali variabili sono in relazione di dipendenza con la quantità d'interesse ed indagare eventuali differenze della struttura dei network bayesiani all'interno della stessa area utilizzando i due differenti algoritmi.

Le relazioni presenti all'interno dei grafi, identificate dagli archi, non vengono qui considerate come relazione di causa ed effetto anche se in alcuni casi può essere plausibile ipotizzare una relazione di questo tipo. L'arco viene interpretato come una relazione di dipendenza diretta tra le due variabili connesse, mentre l'assenza dell'arco implica l'esistenza di una relazione di indipendenza condizionata.

5.3.1 Area geografica: Afro

L'area geografica Afro racchiude tutti i paesi in esame appartenenti al continente africano.

Stato di salute

Le 39702 unità statistiche appartenenti al dataset, che mette in relazione la percezione dello stato di salute con le variabili socio-demografiche e quelle relative allo stato di salute, sono state analizzate dal software Weka ed hanno dato luogo alla seguente struttura:

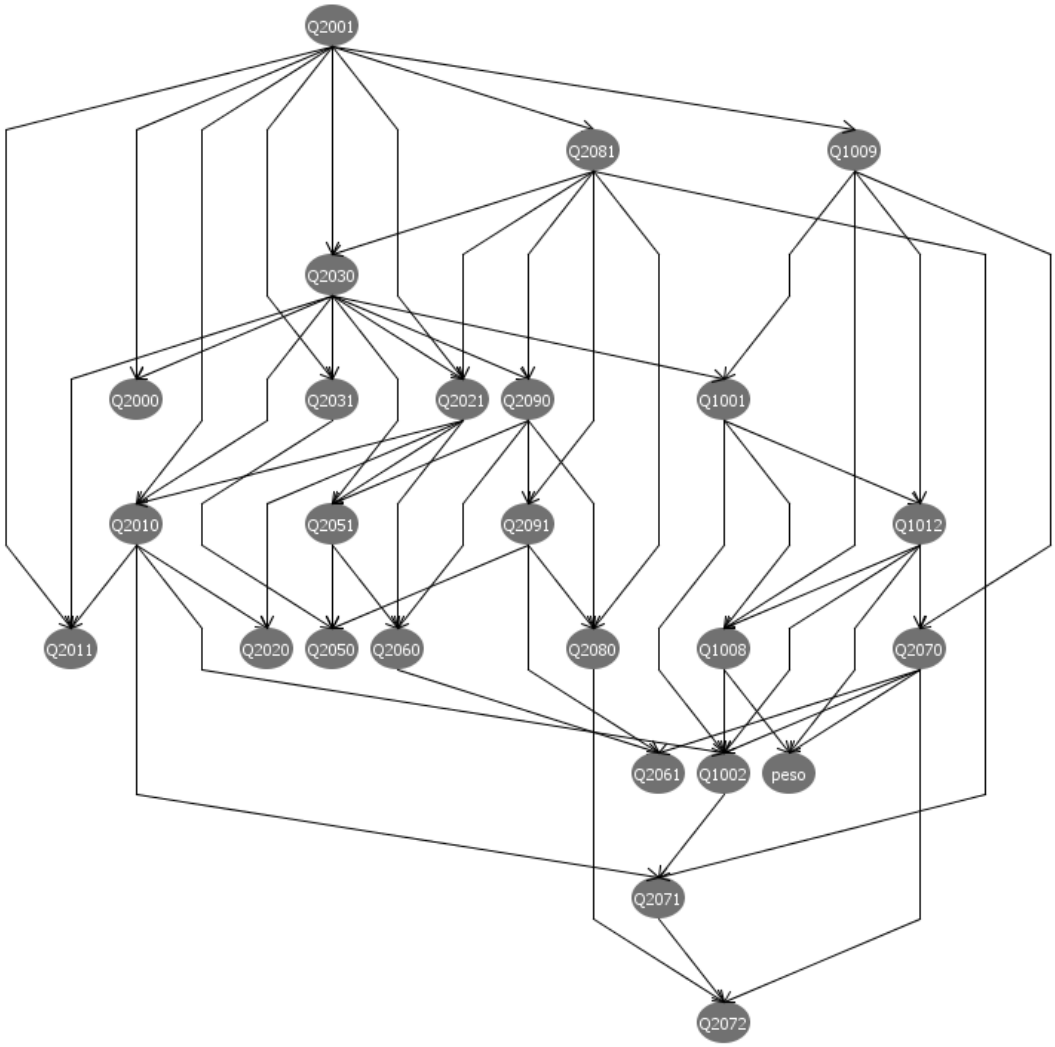


FIGURA 11:STRUTTURA APPRESA, ALGORITMO K2,DATASET STATO DI SALUTE, AREA AFRO

Obbiettivo è valutare la struttura derivante dall'apprendimento dei due differenti algoritmi per indagare le dipendenze esistenti con la variabile d'interesse. Per questo motivo si considera il markov blanket della variabile della percezione dello stato di salute; si riportano inoltre i genitori dei genitori del nodo in esame per poter evidenziare tutte le possibili dipendenze dirette e non. Questi ultimi sono identificati da una linea tratteggiata.

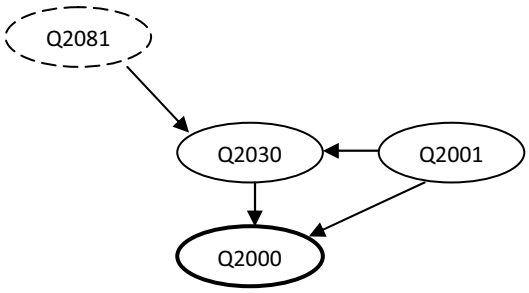


FIGURA 12:INSIEME NODI, ALGORITMO K2,DATASET STATO DI SALUTE, AREA AFRO

Dal MB(Q2000)[30] si deduce che esiste una dipendenza diretta con la difficoltà nello svolgere attività lavorative (Q2001) e con la quantità di dolori o fastidi fisici (Q2030). Esiste una dipendenza non diretta dalla sensazione di stanchezza (Q2081).

Analizzando lo stesso insieme di dati con il software ed omonimo algoritmo BNPC si ottiene la seguente struttura:

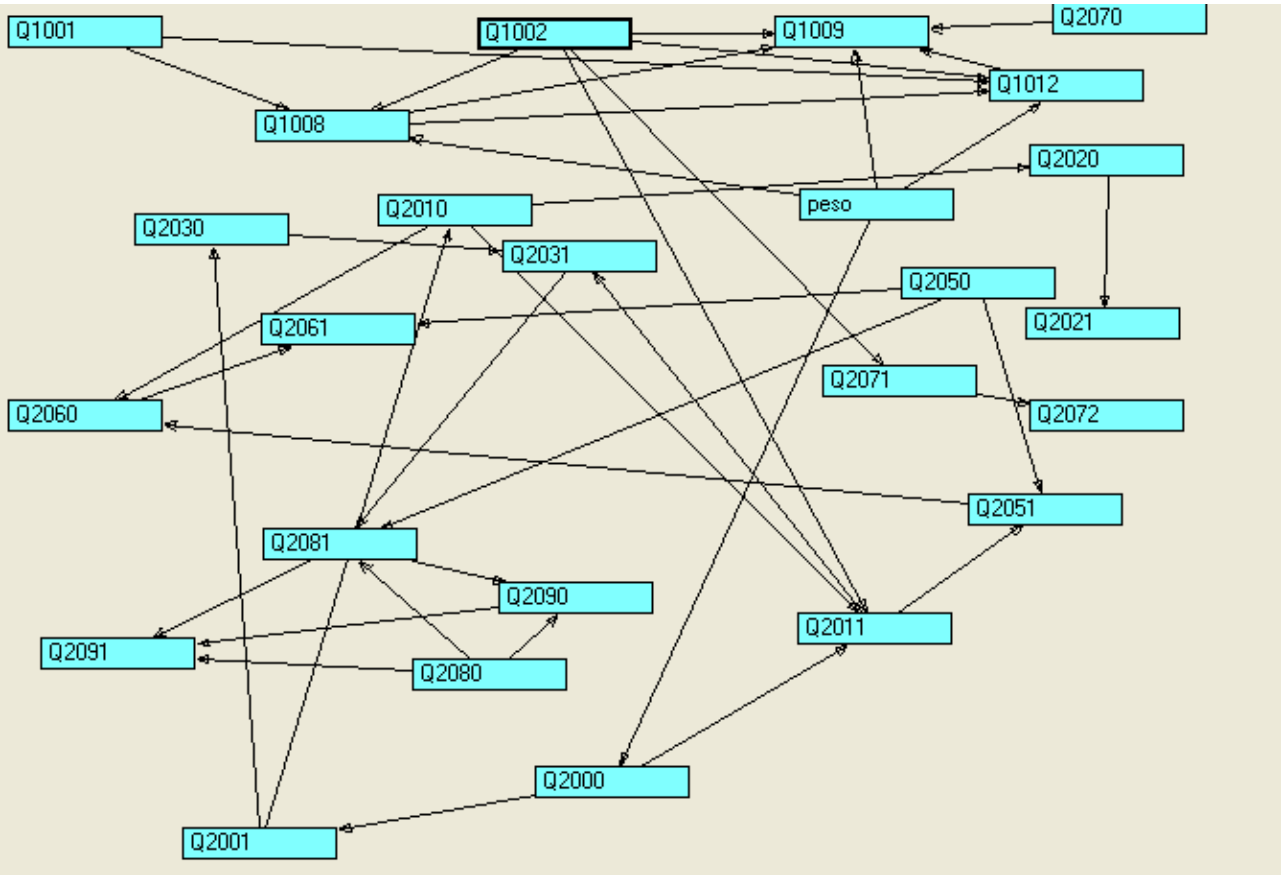


FIGURA 13:STRUTTURA APPRESA, ALGORITMO BNPC,DATASET STATO DI SALUTE, AREA AFRO

Il MB(Q2000) del network bayesiano riportato è:

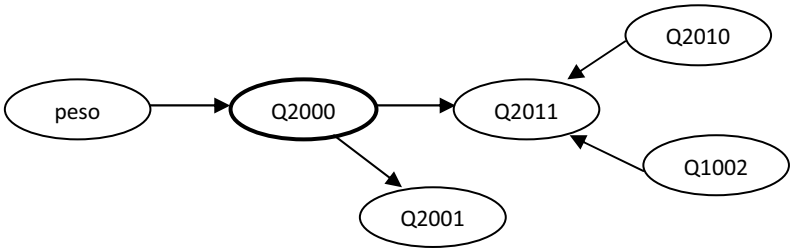


FIGURA 14:INSIEME NODI, ALGORITMO BNPC,DATASET STATO DI SALUTE, AREA AFRO

[30] Sta ad indicare il markov blanket del nodo Q2000.

L'unico genitore della percezione sullo stato di salute individuale è il peso; la variabile Q2000 influenza a sua volta la difficoltà attività lavorativa (Q2001) e la difficoltà nell'effettuare attività vigorose (Q2011). Esistono anche in questo caso delle relazioni di dipendenza con la variabile che descrive la difficoltà nei movimenti (Q2010). E' presente una relazione tra la difficoltà nell'effettuare attività vigorose e la classe d'età dell'unità statistica (Q1002).

I due network bayesiani, derivanti dai diversi metodi di apprendimento sono differenti, l'unico nodo comune ad entrambi risulta essere la variabile relativa alla difficoltà nello svolgere attività lavorative. La prima rete non include alcuna variabile socio-demografica, la seconda prende in considerazione la variabile età.

Fattori di rischio

Si analizza ora il dataset in cui vengono presi in considerazione i fattori di rischio; il campione è composto da 16383 osservazioni:

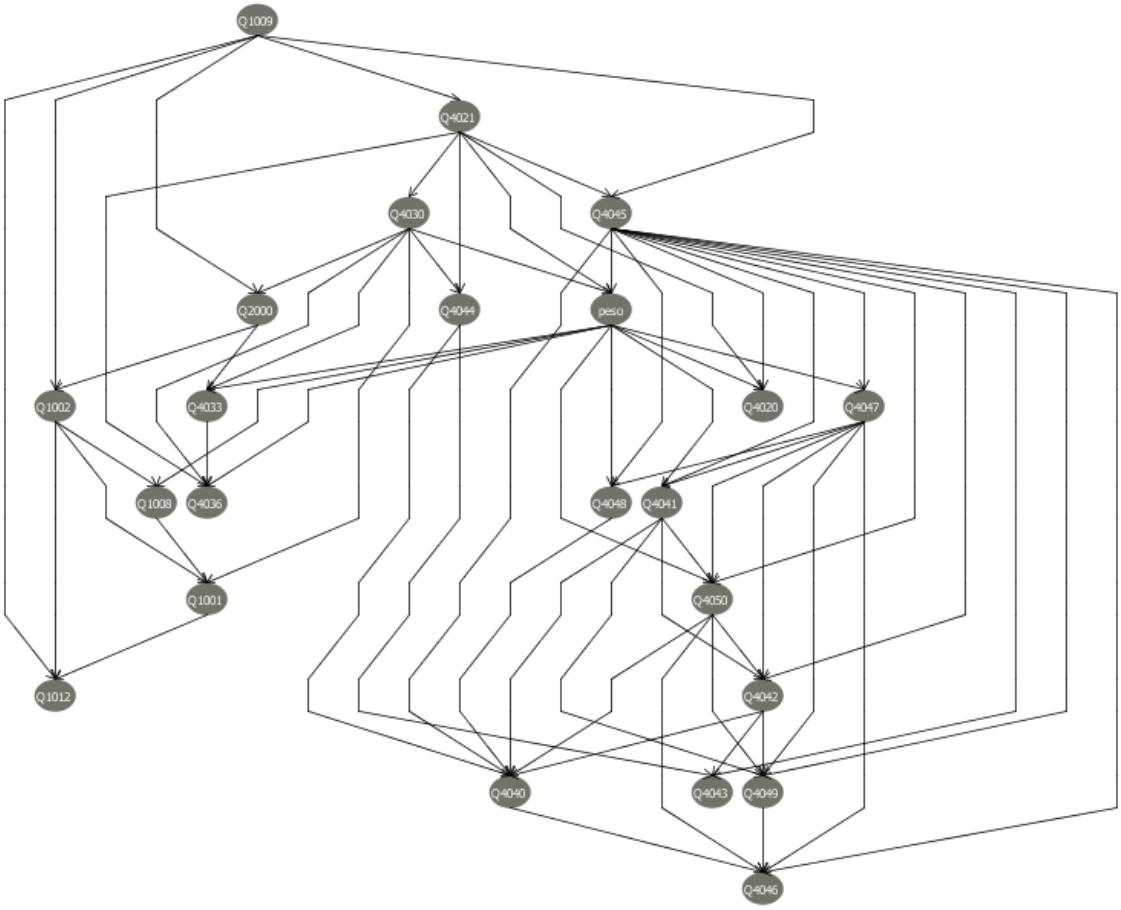


FIGURA 15:STRUTTURA APPRESA, ALGORITMO K2,DATASET FATTORI DI RISCHIO, AREA AFRO

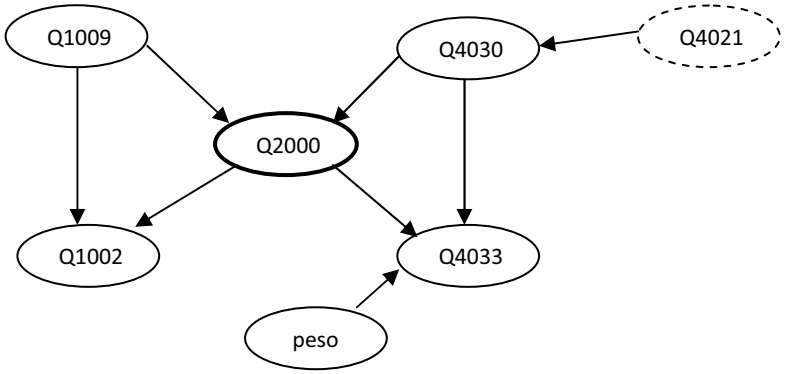


FIGURA 16:INSIEME NODI, ALGORITMO K2,DATASET STATO DI SALUTE, AREA AFRO

Il MB(Q2000) è formato da sei variabili. La percezione dello stato di salute individuale è influenzato direttamente dal livello di educazione (Q1009) e dal numero di giorni in cui si svolgono delle attività fisiche vigorose (Q4030). Inoltre la percezione dello stato di salute è connesso in relazione di dipendenza diretta con la classe d'età (Q1002) ed il numero di giorni in cui vengono effettuate attività fisiche moderate (Q4033). La variabile peso fa parte del markov blanket, anche se non connessa direttamente con la variabile d'interesse. Il numero di vegetali consumati al giorno (Q4021) che non appartiene al markov blanket è però in relazione con la percezione del livello di salute tramite la variabile Q4030.

Tramite l'utilizzo dell'algoritmo BNPC è possibile apprendere la seguente struttura:

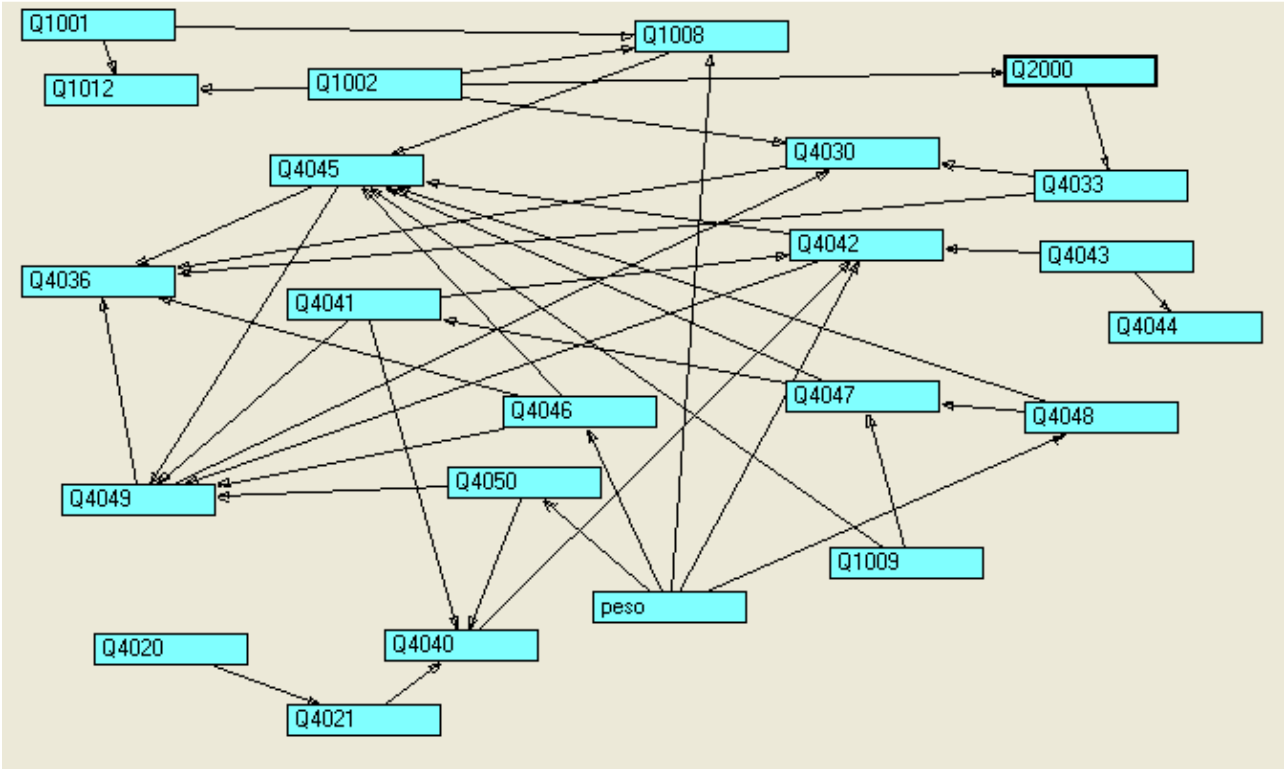


FIGURA 17:STRUTTURA APPRESA, ALGORITMO BNPC,DATASET FATTORI DI RISCHIO, AREA AFRO

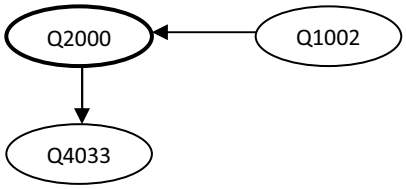


FIGURA 18:INSIEME NODI, ALGORITMO K2,DATASET FATTORI DI RISCHIO, AREA AFRO

Il markov blanket deducibile dal network bayesiano appreso dall'algoritmo BNPC differisce dal precedente. In questo caso la variabile relativa all'età (Q1002) è il genitore della variabile d'interesse che a sua volta è genitore della variabile Q4033 che identifica il numero di giorni in cui vengono svolte le attività moderate.

Confrontando i markov blanket derivanti dalle due strutture apprese, essi risultano essere differenti; entrambi i MB(Q2000) mettono in relazione di dipendenza la variabile percezione dello stato di salute con la variabile età e con il numero di giorni in cui vengono svolte delle attività moderate. Analizzando la struttura derivante dall'algoritmo K2 vengono però considerate anche altre variabili quali il peso, il livello di educazione e, per quanto riguarda i fattori di rischio, il numero di giorni in cui vengono svolte attività vigorose, non solo moderate.

5.3.2 AREA GEOGRAFICA: AMRO

L'insieme di dati qui considerato analizza i dati relativi ai paesi appartenenti all'America che sono stati introdotti nello studio.

Stato di salute

Il sotto-campione in esame conta 44652 unità statistiche. Anche in questo caso si considerano le variabili socio-demografiche e le variabili riguardanti lo stato di salute.

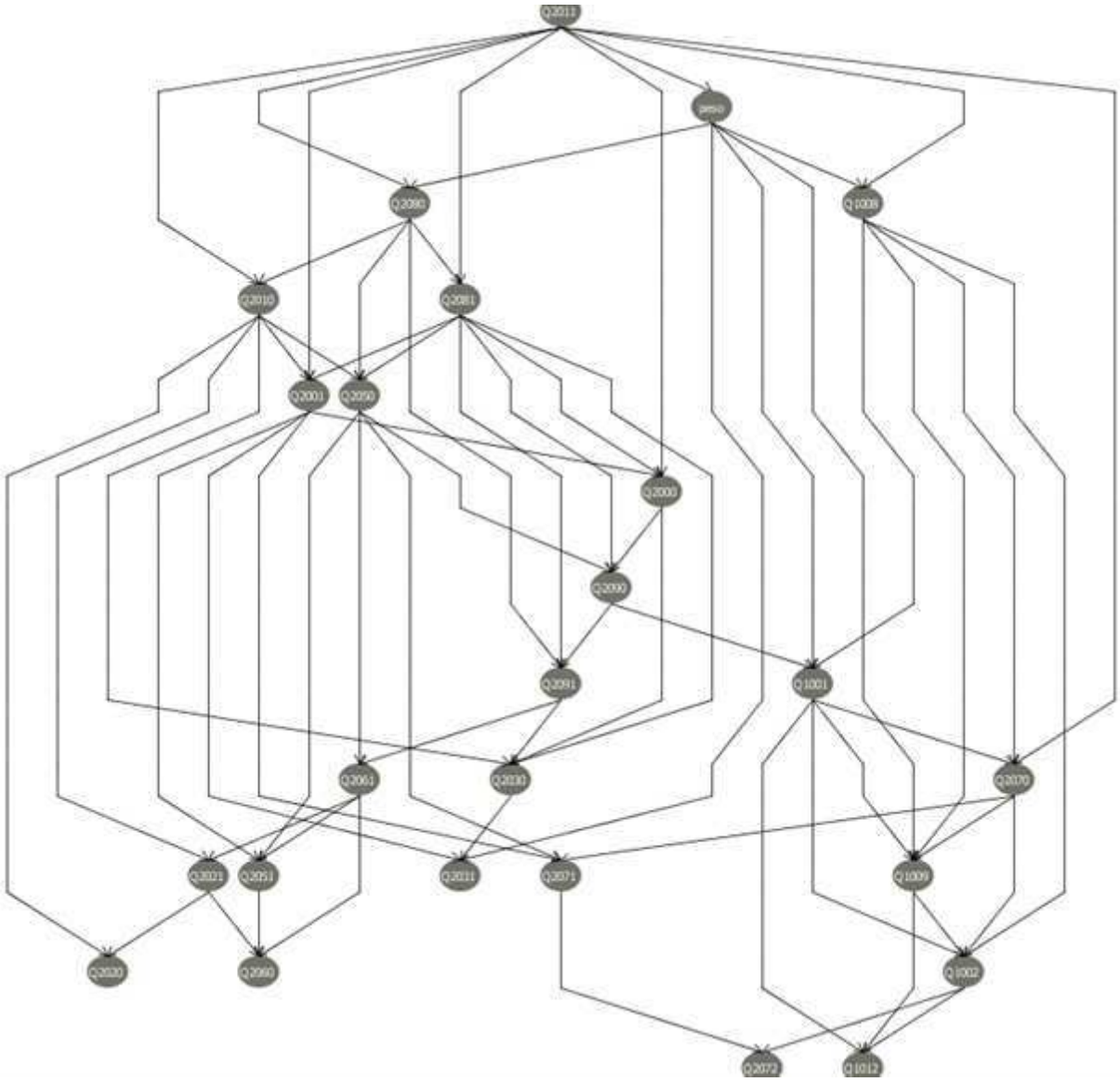


FIGURA 19:STRUTTURA APPRESA, ALGORITMO K2,DATASET STATO DI SALUTE, AREA AMRO

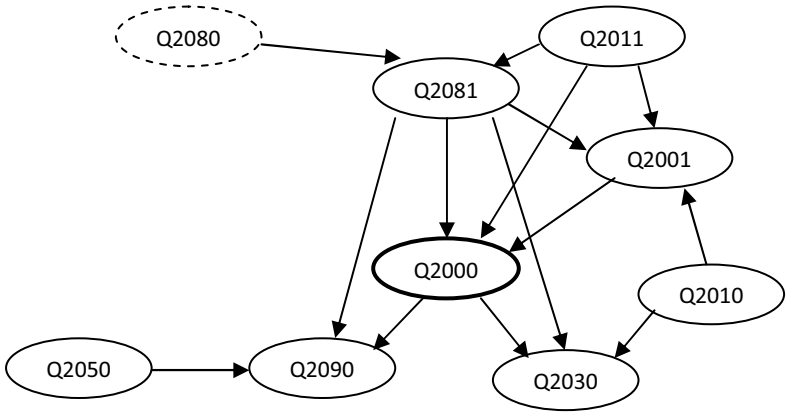


FIGURA 20:INSIEME NODI, ALGORITMO K2,DATASET STATO DI SALUTE, AREA AMRO

L'insieme di variabili riportato include nove variabile di cui sette costituiscono il MB(Q2000): la percezione dello stato di salute individuale è in relazione di dipendenza diretta con la variabile che descrive la difficoltà dei movimenti vigorosi (Q2011), con la sensazione di stanchezza (Q2081) e con la difficoltà nello svolgere attività lavorative (Q2001); inoltre esistono una dipendenza diretta con la variabile che descrive la sensazione di tristezza o depressione (Q2090) e un'altra con la quantità di doloro fisici riscontrati (Q2030).
Del MB(Q2000) fanno parte altre tre variabili: avere difficoltà nei movimenti (Q2010) e la difficoltà nel ricordare concetti o concentrarsi (Q2050). La percezione dello stato di salute dipende, anche se non direttamente, dai problemi di insonnia (Q2080).

Si apprende ora dallo stesso insieme di dati la struttura del grafo secondo un approccio constraint-based:

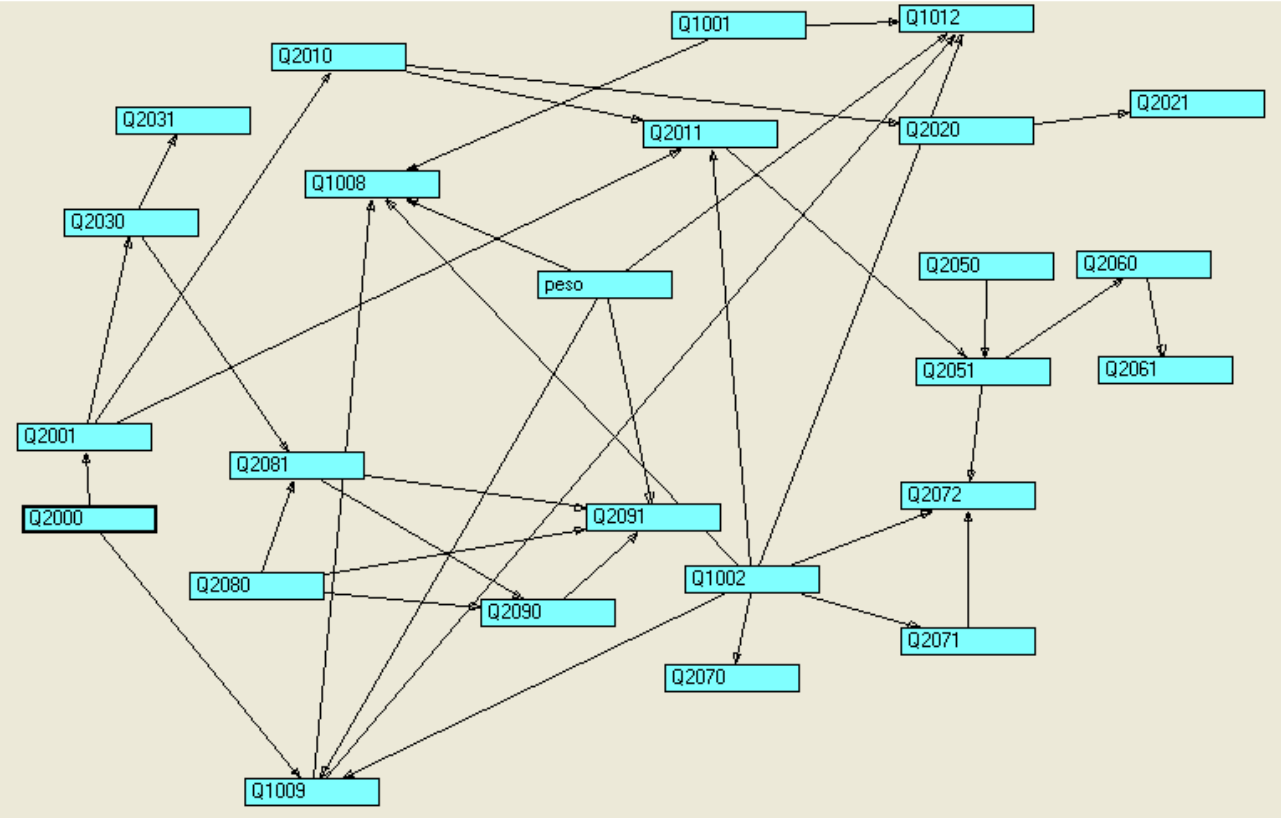


FIGURA 21:STRUTTURA APPRESA, ALGORITMO BNPC,DATASET STATO DI SALUTE, AREA AMRO

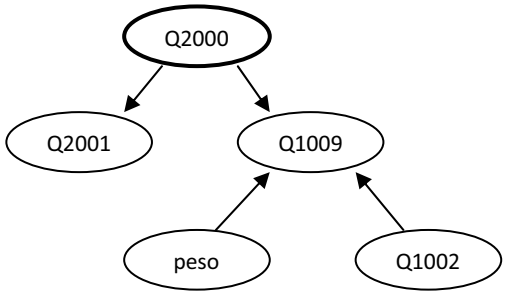


FIGURA 22:INSIEME NODI, ALGORITMO BNPC,DATASET STATO DI SALUTE, AREA AMRO

A differenza delle situazioni sin qui riportate, i due markov blanket derivanti dall'apprendimento della struttura mediante un algoritmo BNPC sono totalmente differenti. In questo caso le variabili connesse direttamente con Q2000 sono le variabili che descrivono la difficoltà nello svolgere attività lavorative (Q2001) ed il livello di educazione di un individuo (Q1009). All'interno dell'insieme di nodi considerato sono presenti anche il peso e l'età dell'individuo (Q1002).
L'unica variabile che in entrambi i casi risulta essere dipendente è la difficoltà nello svolgere attività lavorative.

Da notare che la struttura derivante dall'analisi dell'algoritmo K2 mette in evidenza che la percezione dello stato di salute non dipende dalle variabili socio-demografiche, ma esclusivamente dalle variabili relative allo stato di salute. Al contrario il secondo network costruito sembra dipendere esclusivamente da variabili socio-demografiche, fatta esclusione della variabile che descrive la difficoltà nello svolgere attività lavorative.

FATTORI DI RISCHIO

L'insieme delle osservazioni appartenenti ai paesi dell'America in cui si considerano i fattori di rischio sono 2668.

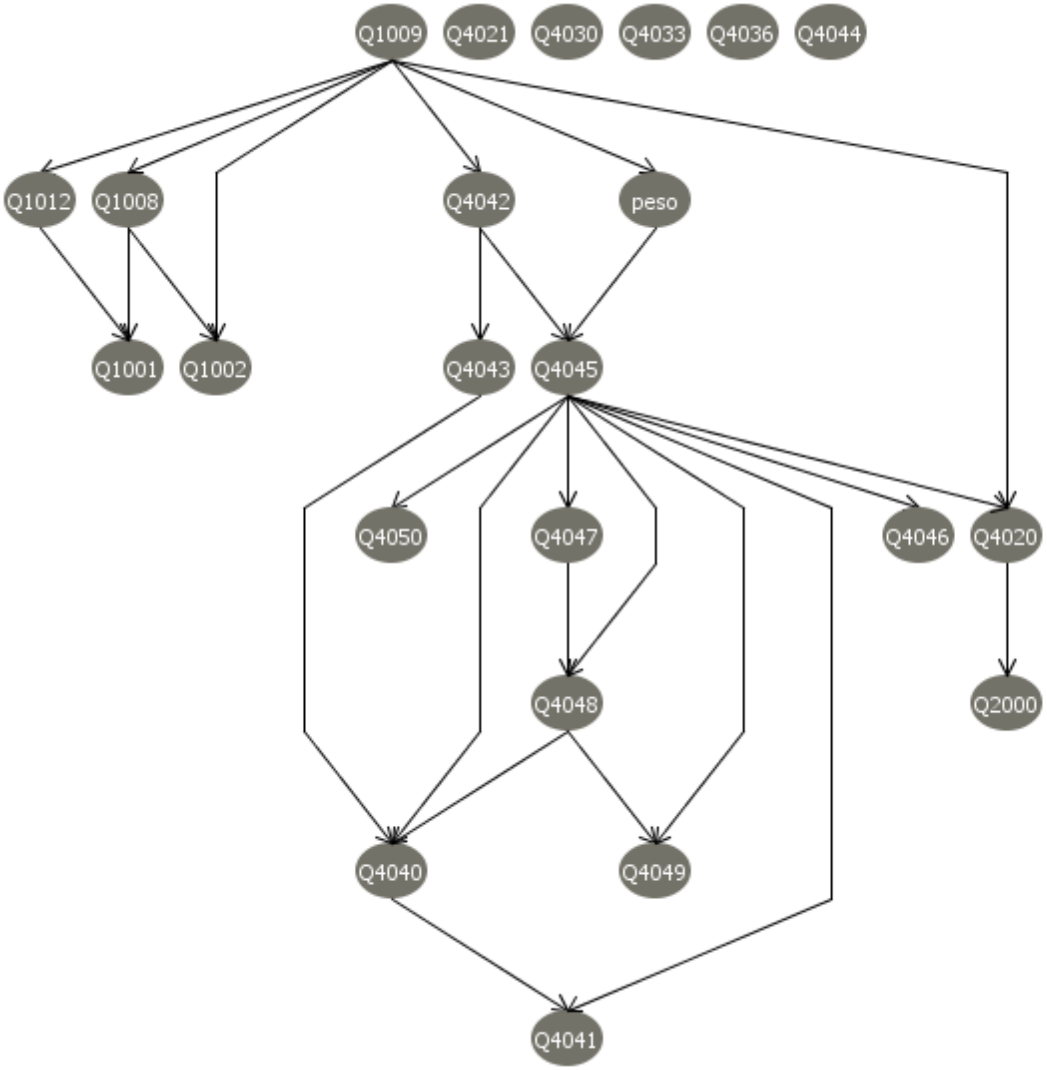


FIGURA 23:STRUTTURA APPRESA, ALGORITMO K2,DATASET FATTORI DI RISCHIO, AREA AMRO

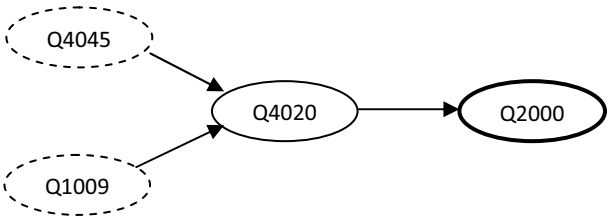


FIGURA 24:INSIEME NODI, ALGORITMO K2,DATASET FATTORI DI RISCHIO, AREA AMRO

Analizzando il markov blanket di Q2000, si evince che l'unico nodo che si trova in relazione di dipendenza con la percezione di salute individuale riguarda la variabile che descrive le abitudini alimentari degli individui intervistati, in particolare Q4020 indica quanti frutti vengono consumati ogni giorno. Quest'ultima dipende dal livello di educazione e dal tipo di servizi igienici sono presenti nell'abitazione.

Il markov blanket in questo caso può essere considerato poco consueto; risulta interessante analizzare la struttura riportata di seguito costruita mediante l'ausilio dell'algoritmo BNPC.

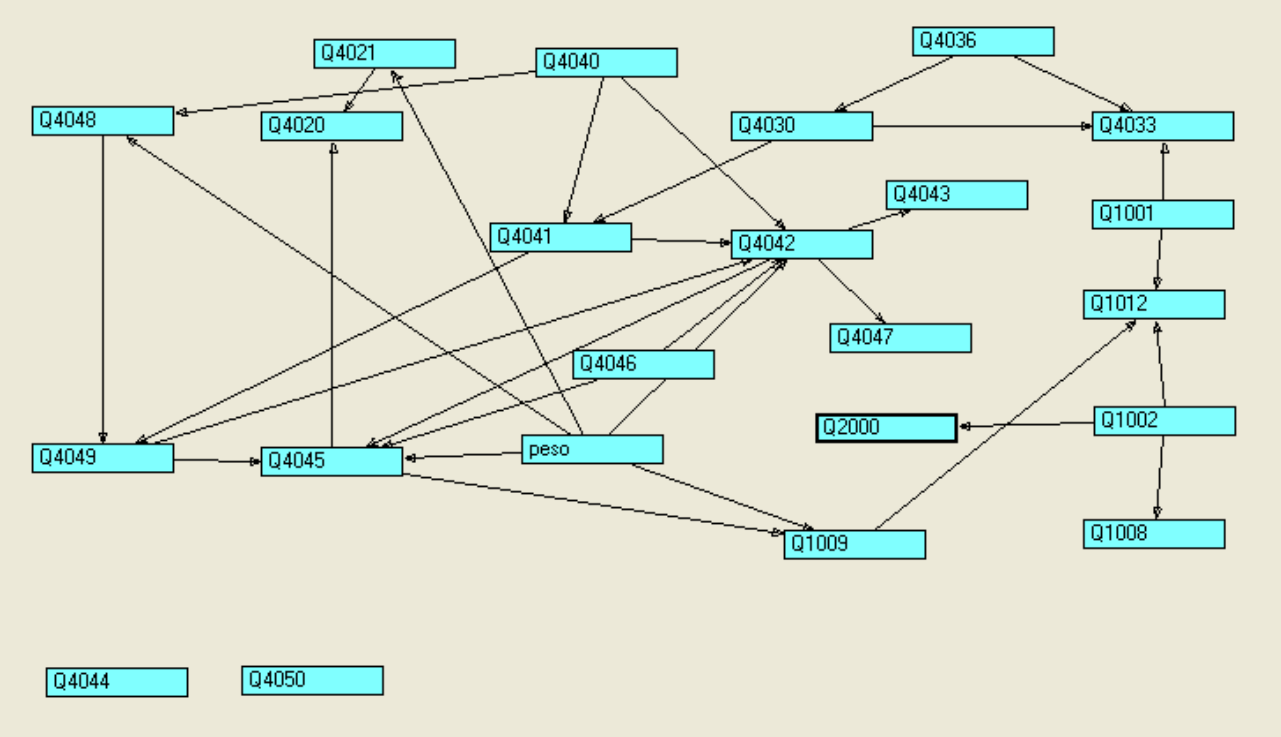


FIGURA 25:STRUTTURA APPRESA, ALGORITMO BNPC,DATASET FATTORI DI RISCHIO, AREA AMRO

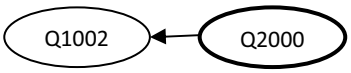


FIGURA 26:INSIEME NODI, ALGORITMO BNPC,DATASET FATTORI DI RISCHIO, AREA AMRO

(Q1002) è l'unica variabile per cui è possibile notare una relazione di dipendenza.

Un aspetto rilevante è che sia nella struttura appresa dall'algoritmo Weka che dall'algoritmo BNPC esiste solo una relazione di dipendenza diretta.

Inoltre in entrambe le sotto-reti riportate esistono delle variabili per cui non è possibile rintracciare alcuna relazione di dipendenza. Per la struttura appresa dall'algoritmo K2 si ha che Q4021(numero di vegetali consumati al giorno), Q4030(Numero di giorni in cui si praticano attività vigorose), Q4033(Numero di giorni in cui si praticano attività moderate), Q4036(Numero di giorni in cui si praticano camminate), Q4044(Possibilità di utilizzare venti litri di acqua al giorno).

Per la struttura appresa dall'algoritmo BNPC solo due variabili non risultano essere indipendenti dal resto dell'insieme di variabili, si tratta Q4044(Possibilità di utilizzare venti litri di acqua al giorno) e Q4050(Possibilità di riscaldare la casa).

5.3.3 AREA GEOGRAFICA: EMRO

Il campione derivante delle interviste del Mediterraneo Orientale conta 11345 osservazioni.

Stato di salute

Si analizza di seguito la struttura derivante dall'analisi delle variabili socio-demografiche e quelle inerenti lo stato di salute:

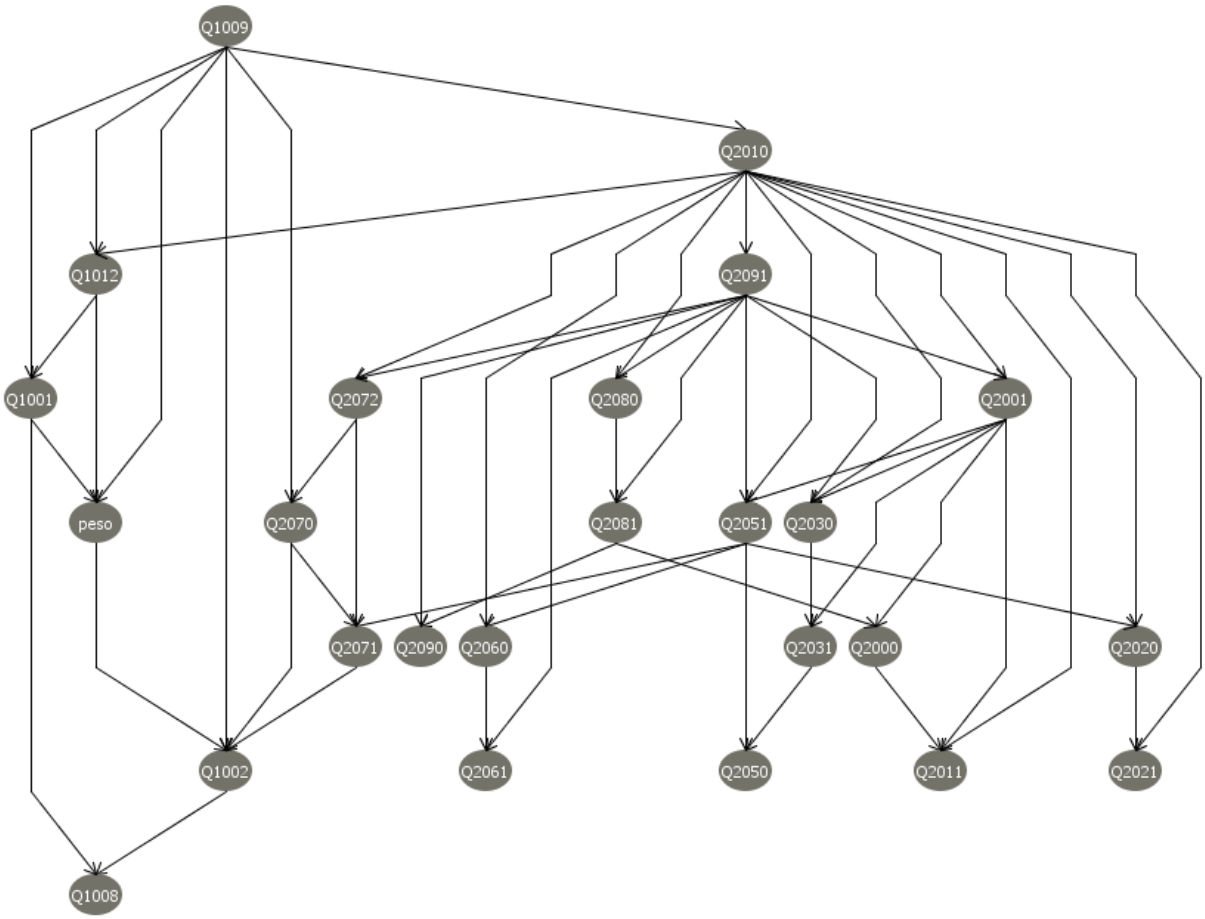


FIGURA 27:STRUTTURA APPRESA, ALGORITMO K2, DATASET STATO DI SALUTE, AREA EMRO

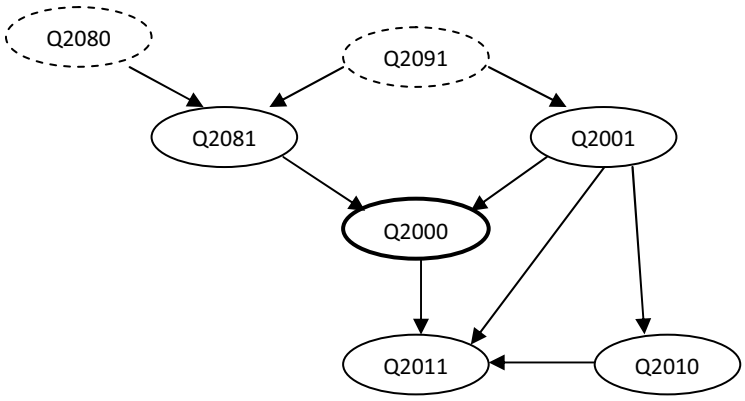


FIGURA 28:INSIEME NODI, ALGORITMO K2, DATASET STATO DI SALUTE, AREA EMRO

La percezione dello stato di salute individuale si trova in dipendenza diretta con la sensazione di stanchezza (Q2081), con la difficoltà nello svolgere attività lavorative (Q2001) e con la difficoltà nello svolgere attività vigorose (Q2011). E' presente all'interno del markov blanket di Q2000 anche la variabile che descrive la difficoltà dei movimenti (Q2010). L'insieme riportato include anche altre variabile in relazione di dipendenza con la variabile d'interesse: problemi di insonnia (Q2080), sensazione di tristezza o depressione (Q2090).

La struttura sul medesimo insieme di dati appreso dall'algoritmo BNPC è la seguente:

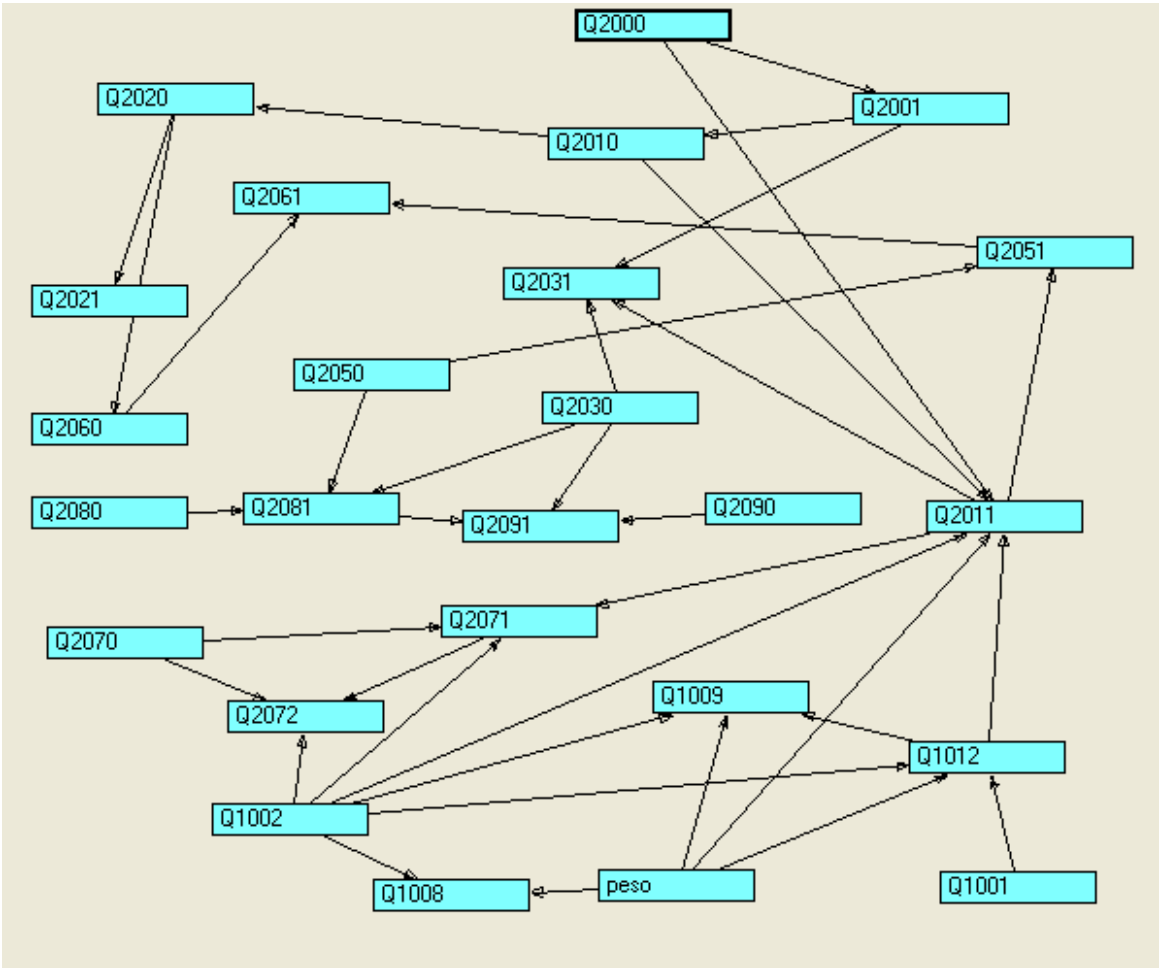


FIGURA 29:STRUTTURA APPRESA, ALGORITMO BNPC, DATASET STATO DI SALUTE, AREA EMRO

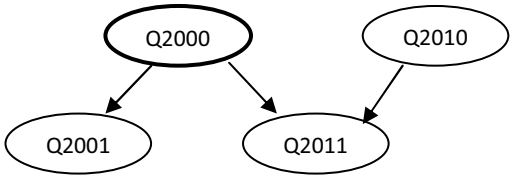


FIGURA 30:INSIEME NODI, ALGORITMO BNPC, DATASET STATO DI SALUTE, AREA EMRO

Tutte le variabili prese in esame in questa sotto-struttura si ritrovano nella struttura appresa mediante l'algoritmo K2. Non tutte le variabili considerate sono in relazione di dipendenza diretta come nel caso precedente, infatti la difficoltà nei movimenti (Q2010) non ha una dipendenza diretta. I due markov blanket evidenziano, escludendo la sensazione di stanchezza, la stessa struttura di dipendenza per la variabile Q2000.

Entrambi gli algoritmi mettono in relazione le stesse variabili; si può quindi affermare che per i paesi del Mediterraneo Orientale la percezione dello stato di salute è influenzata dalle difficoltà nello svolgere attività lavorative, la difficoltà nel muoversi e la difficoltà nel compiere attività vigorose.
L'algoritmo K2 include altre relazioni di dipendenza qui non presenti: sensazione di stanchezza (Q2081), problemi di insonnia (Q2080) e sensazione di tristezza o depressione (Q2090).

Fattori di rischio

L'insieme di dati riguardanti i fattori di rischio è composto da 2616 unità statistiche. La struttura appresa mediante il software Weka è la seguente:

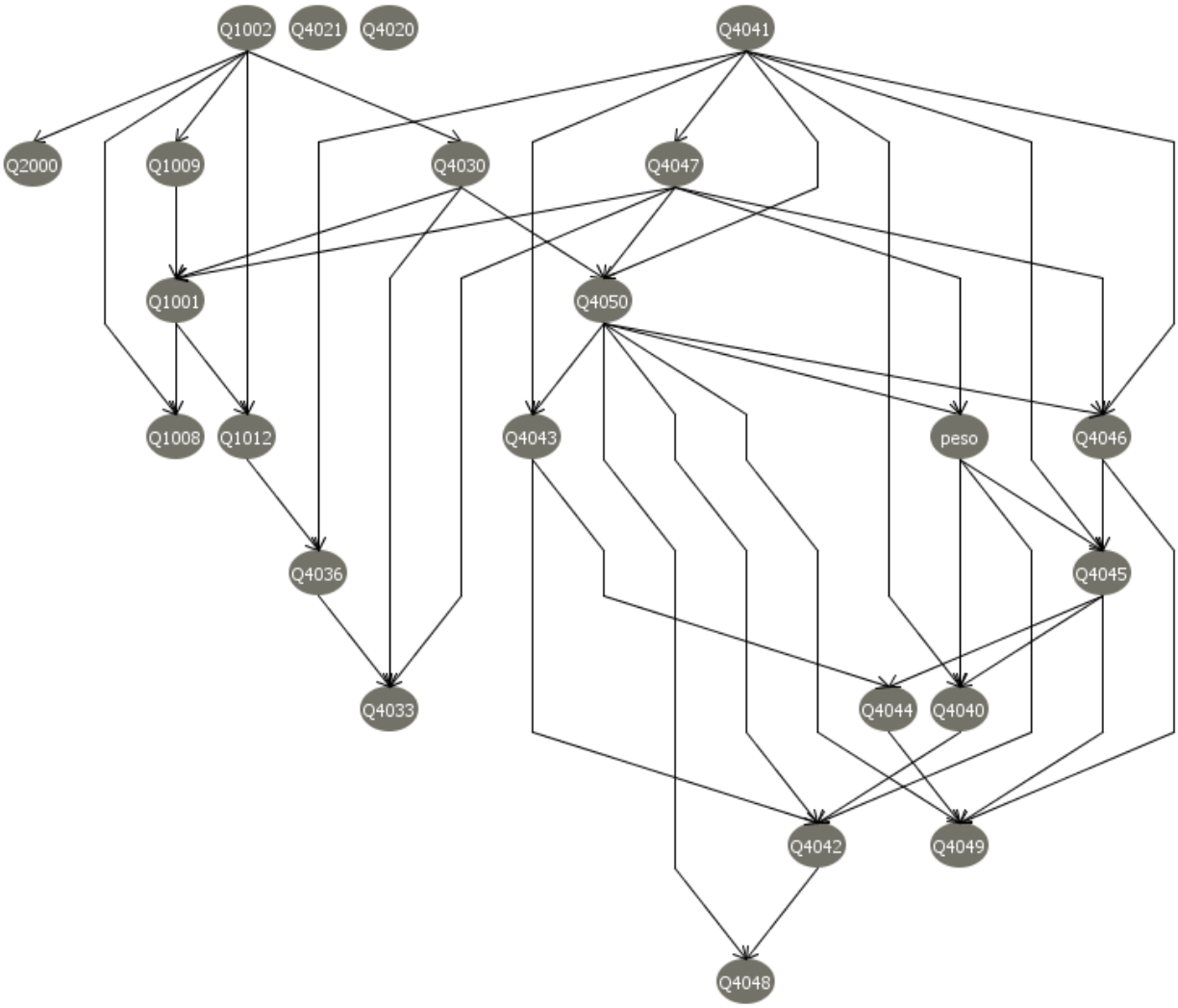


FIGURA 31:STRUTTURA APPRESA, ALGORITMO K2,DATASET FATTORI DI RISCHIO, AREA EMRO

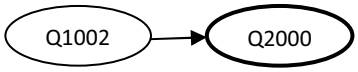


FIGURA 32:INSIEME NODI, ALGORITMO K2,DATASET FATTORI DI RISCHIO, AREA EMRO

L'unico nodo che è in relazione con Q2000 è una variabile appartenente all'insieme delle variabili socio-demografiche: Q1002, che descrive l'età dei soggetti.

Si analizza ora il dataset mediante l'algoritmo BNPC:

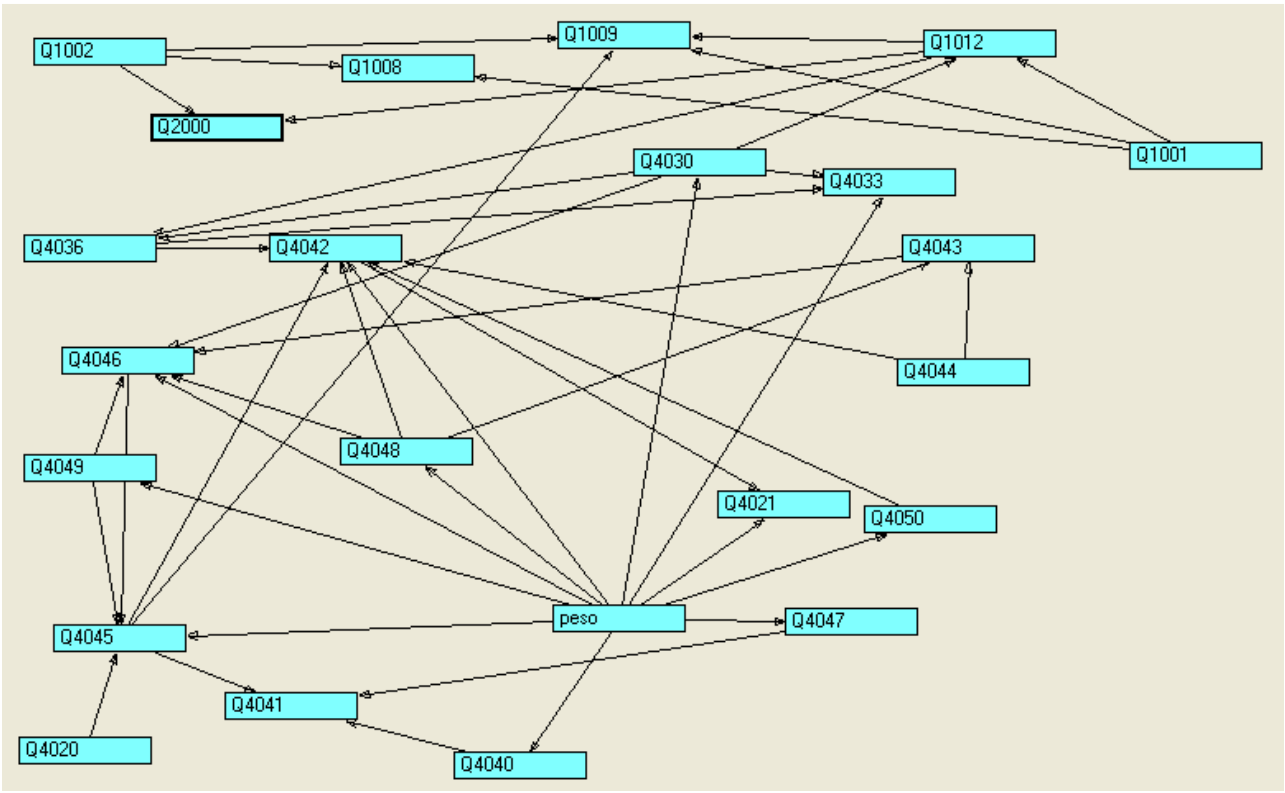


FIGURA 33:STRUTTURA APPRESA, ALGORITMO BNPC,DATASET FATTORI DI RISCHIO, AREA EMRO

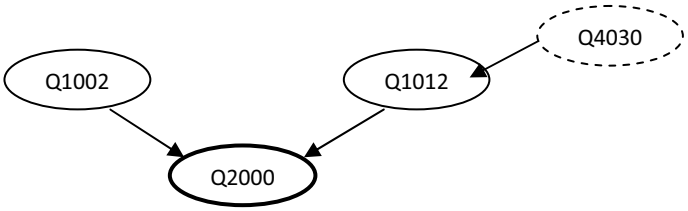


FIGURA 34:INSIEME NODI, ALGORITMO BNPC,DATASET FATTORI DI RISCHIO, AREA EMRO

Quanto descritto dalla precedente struttura, trova conferma in questo markov blanket della variabile d'interesse. Non sono presenti variabili appartenenti all'insieme dei fattori di rischio ma esclusivamente variabili socio-demografiche, in particolare l'età (Q1002) ed il lavoro corrente (Q1012) dell'intervistato.

Analizzando l'intero insieme, non solo il MB(Q2000), si nota una dipendenza indiretta con la variabile che descrive il numero di giorni in cui vengono effettuate attività fisiche vigorose.

Ancora una volta si evince che analizzando esclusivamente il MB della variabile d'interesse, le strutture sono simili; se si allarga l'insieme, ossia prendendo in considerazione i genitori dei genitori della variabile Q2000, le strutture differiscono.
Sembra importante sottolineare che per i cittadini del Mediterraneo Orientale la percezione di salute non dipenda dai fattori di rischio, quanto dalle variabili socio-demografiche.

5.3.4 Area geografica: Euro

L'insieme Euro analizza i paesi appartenenti all'Europa. Come nelle altre sezioni si prende in esame prima l'insieme per valutare le relazioni di interesse tra la percezione di salute e quelle inerenti allo stato di salute; in un secondo momento si analizzerà l'insieme di variabili riguardanti i fattori di rischio.

Stato di salute

Le unità statistiche che compongono l'insieme di interesse sono 30125.

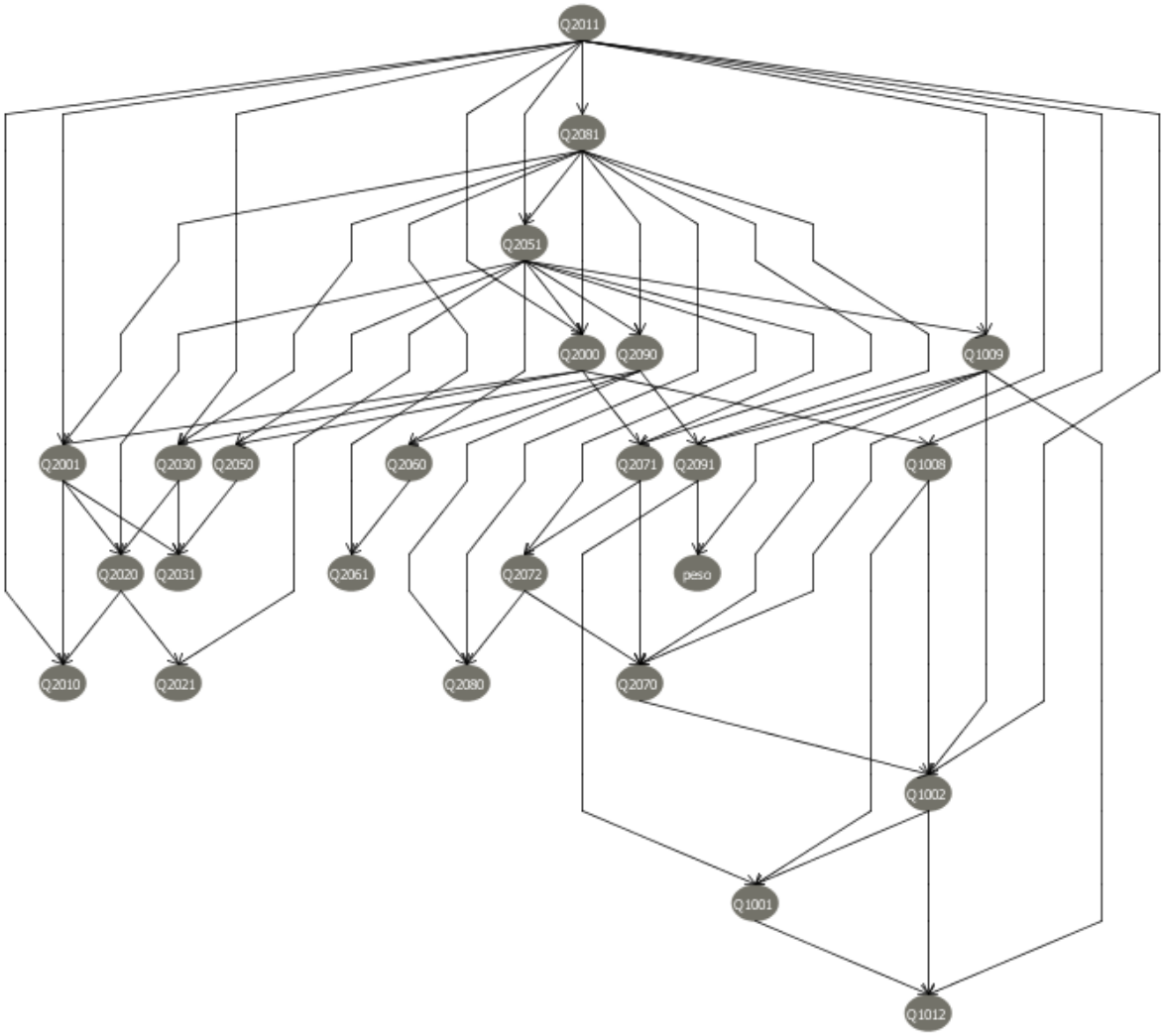


FIGURA 35:STRUTTURA APPRESA, ALGORITMO K2, DATASET STATO DI SALUTE, AREA EURO

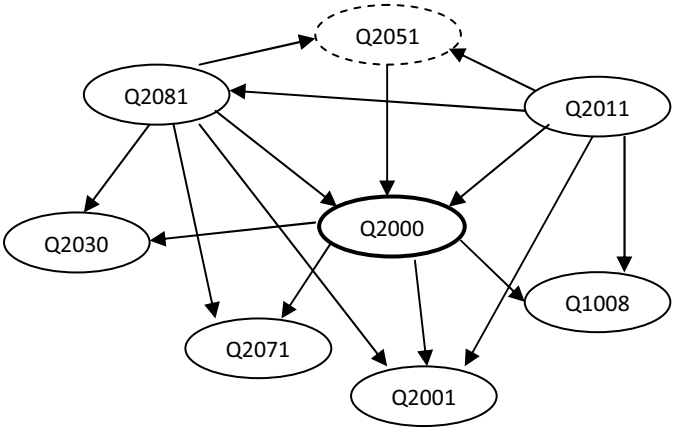


FIGURA 36:INSIEME NODI, ALGORITMO K2,DATASET STATO DI SALUTE, AREA EURO

L'insieme risultante è complesso. Sono implicate dieci variabili e molte dipendenze: di queste quelle dirette con la variabile percezione dello stato di salute sono la difficoltà nello svolgere attività lavorativa (Q2011), la difficoltà ad imparare concetti nuovi (Q2051), il provare sensazioni di stanchezza (Q2081), la quantità di dolori o fastidi fisici (Q2030), la difficoltà nel vedere e riconoscere persone (Q2071), la difficoltà nello svolgere attività lavorativa (Q2001) ed infine lo stato matrimoniale (Q1008).
A differenza di tutte le altre sotto-strutture, in questa tutte le dipendenze presenti sono dirette.

Si fa ora riferimento alla struttura appresa dall'algoritmo BNPC:

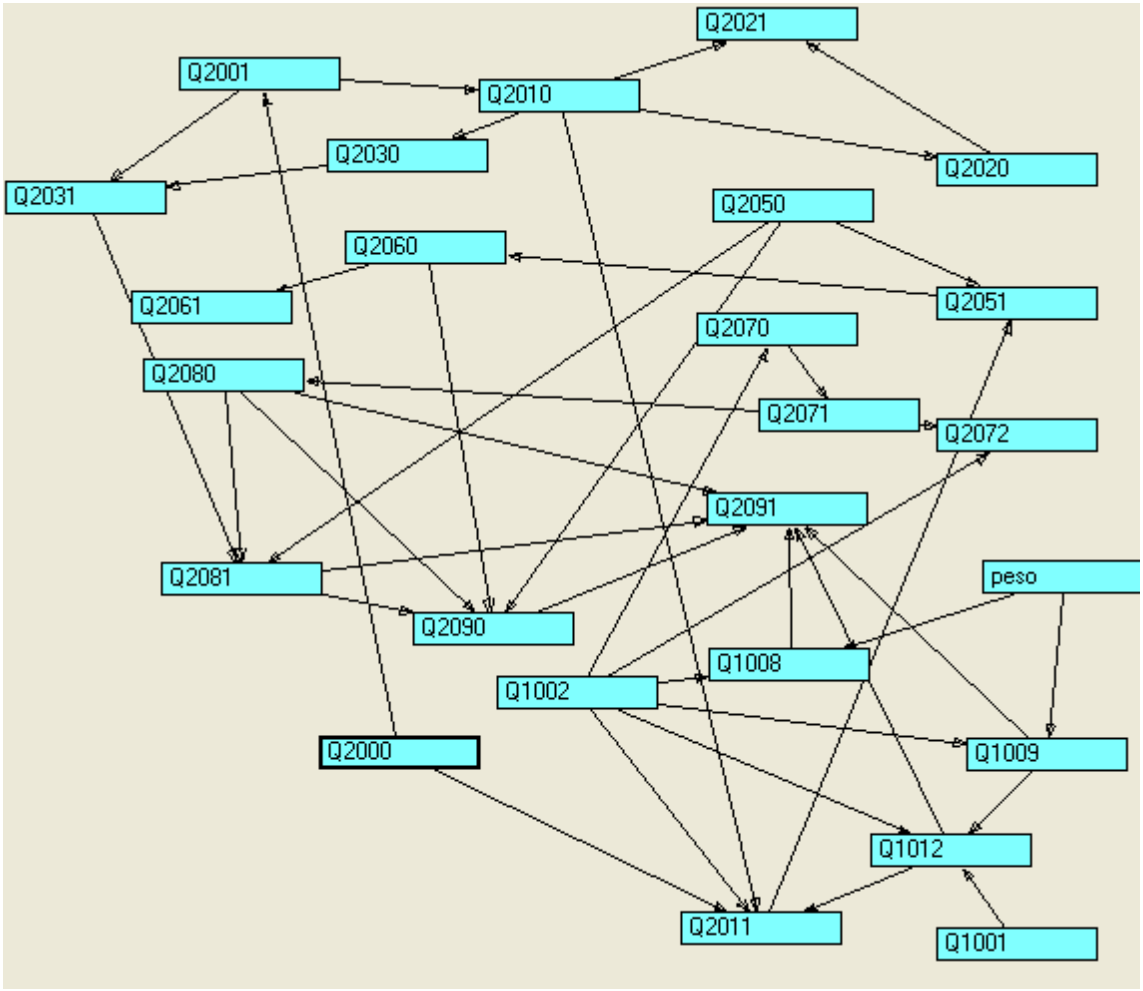


FIGURA 37:STRUTTURA APPRESA, ALGORITMO BNPC,DATASET STATO DI SALUTE, AREA EURO

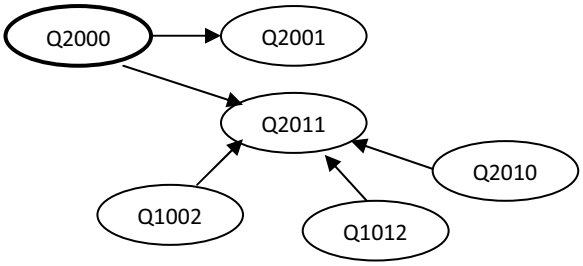


FIGURA 38:INSIEME NODI, ALGORITMO BNPC,DATASET STATO DI SALUTE, AREA EURO

La differenza tra le due strutture è qui più evidente che in altre aree geografiche: le uniche relazioni dirette riguardano le difficoltà nelle attività lavorativa (Q2001) e la difficoltà nell'affrontare attività vigorose (Q1002). Le difficoltà nel muoversi (Q2010) ed il lavoro corrente (Q1012) sono le variabili dipendenti ma non direttamente connesse.

La differenza tra i due markov blanket della percezione della salute dedotti sono molto diversi: nel primo sotto-network bayesiano sono presenti meno archi rispetto alla seconda retta; infatti, nella struttura appresa dall'algoritmo BNPC è presente un numero minore di relazioni di dipendenza rispetto alla quantità esistente delle relazioni nel primo grafo.
Nel primo network bayesiano si ripropone la stessa situazione vista nelle aree geografiche precedenti: le variabili socio-demografiche non ricoprono un ruolo fondamentale per le relazioni di dipendenza con la variabile d'interesse; nel MB(Q2000) sono presenti sia variabili socio-demografiche, che variabili relative lo stato di salute degli intervistati.

Fattori di rischio

Il dataset qui esaminato è composto da 3471 osservazioni. La struttura appresa mediante l'utilizzo dell'algoritmo K2 è la seguente:

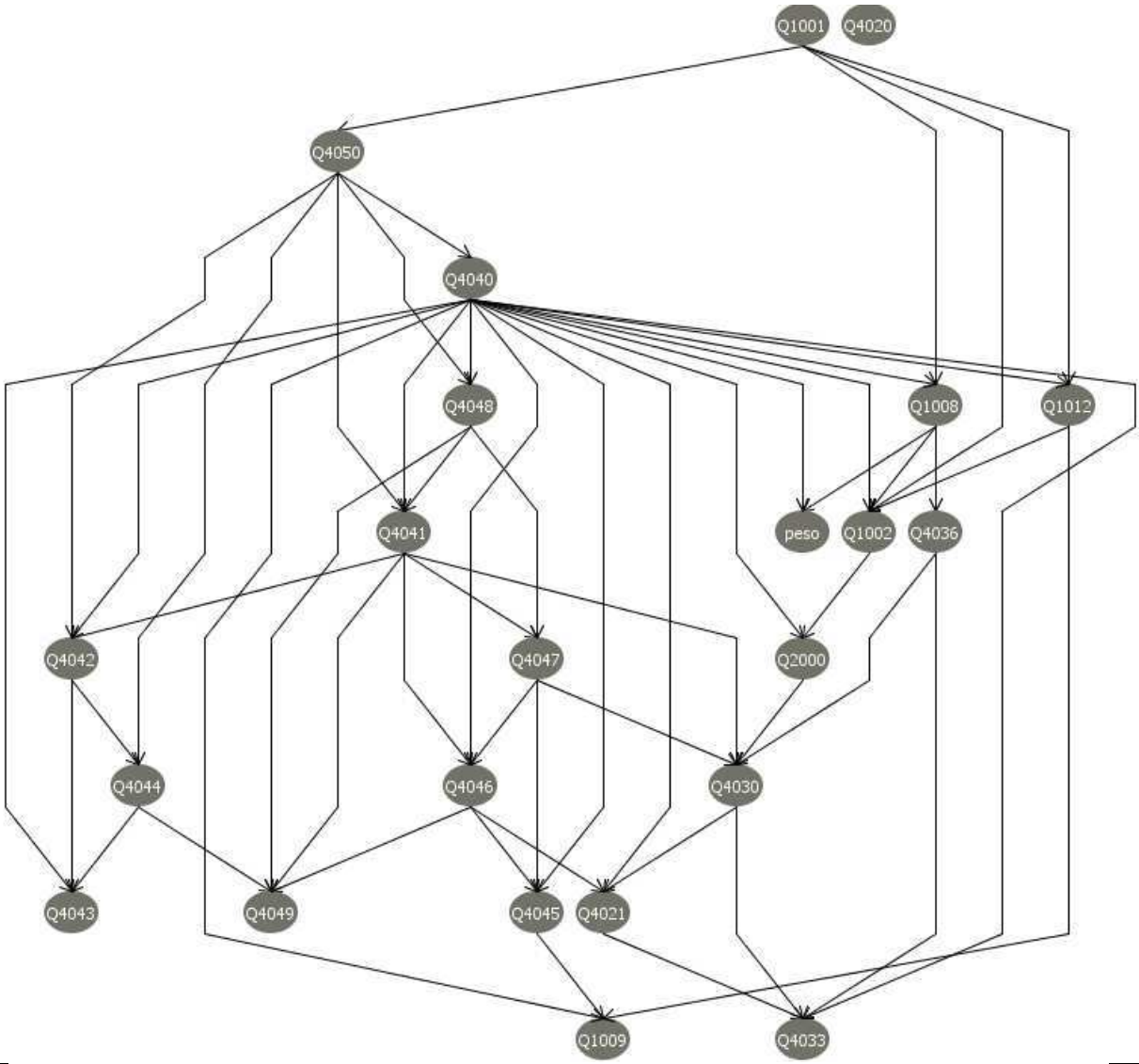


FIGURA 39:STRUTTURA APPRESA, ALGORITMO K2,DATASET FATTORI DI RISCHIO, AREA EURO

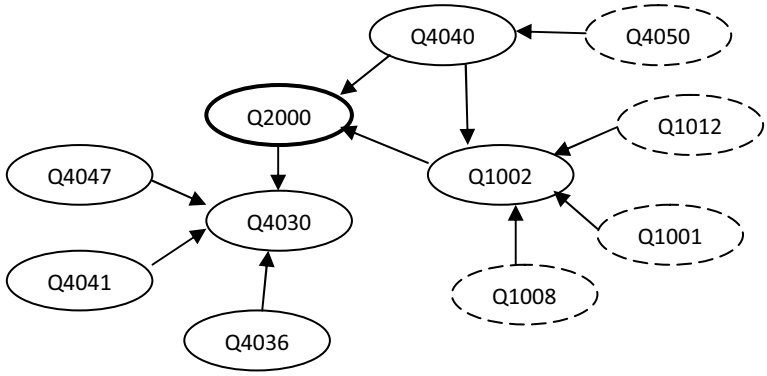


FIGURA 40:INSIEME NODI, ALGORITMO K2,DATASET FATTORI DI RISCHIO, AREA EURO

Le variabili direttamente connesse con Q2000 riguardano l'età del soggetto (Q1002), il tipo di pavimento presente nell'abitazione (Q4040) ed il numero di giorni in cui si effettuano attività vigorose (Q4030). Altri

fattori di rischio appartenenti al markov blanket influenzano la variabile d'interesse: il numero di giorni in cui si effettuano lunghe camminate (Q4036), il tipo di pareti presenti nell'abitazione (Q4041) ed infine il tipo di risorsa che si utilizza per cucinare (Q4047). In questo caso sembra che la percezione dello stato di salute per i paesi europei dipenda maggiormente dalle condizioni in cui gli intervistati vivono nel quotidiano.

La possibilità di riscaldare la casa (Q4050) è in relazione di dipendenza non diretta con la variabile d'interesse; tramite la Q1002 è possibile evidenziare un insieme di variabili socio-demografiche: il sesso dell'intervistato (Q1001), lo stato matrimoniale (Q1008), il lavoro corrente (Q1012).

Sulla base delle informazioni deducibili dallo stesso dataset, si apprende la struttura utilizzando l'algoritmo BNPC:

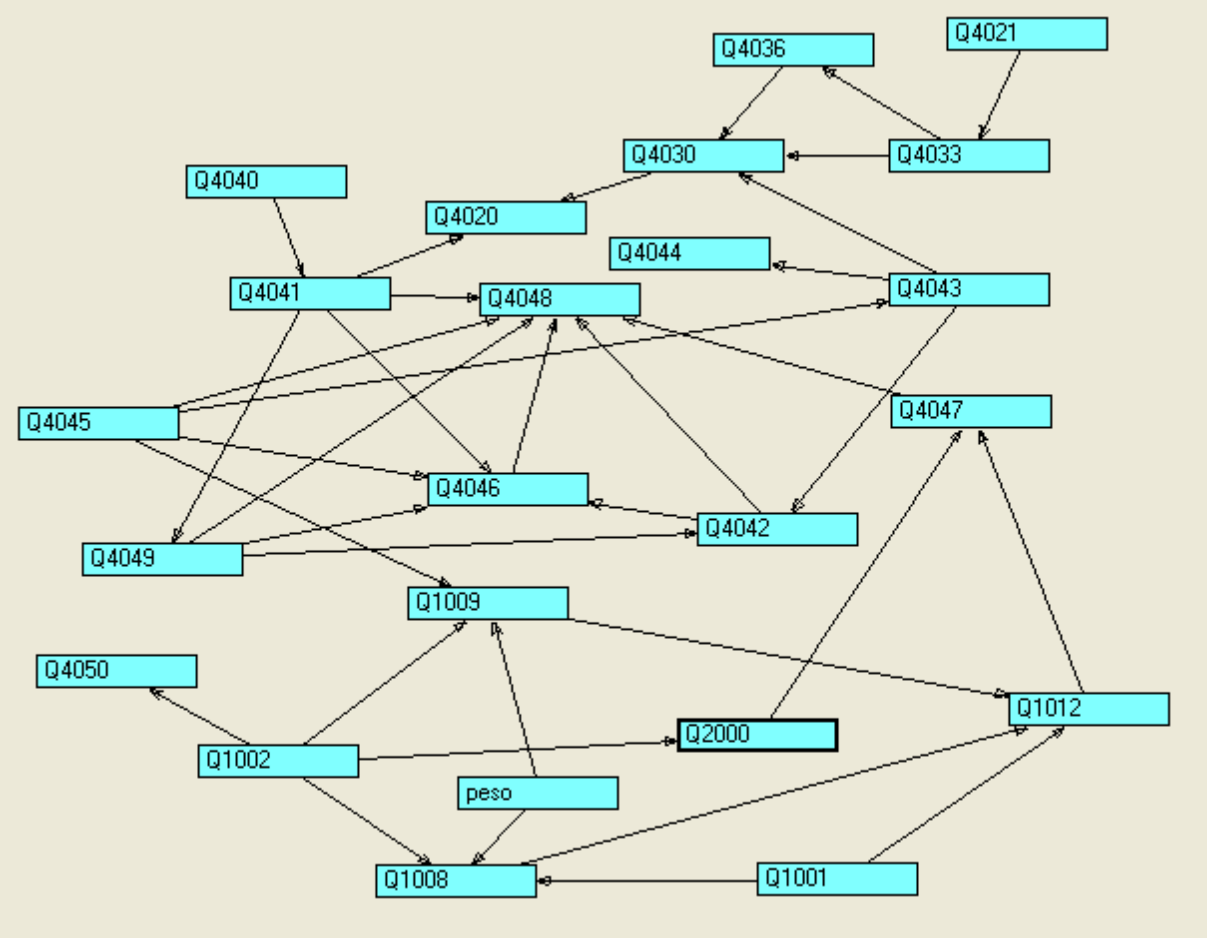


FIGURA 41:STRUTTURA APPRESA, ALGORITMO BNPC, DATASET FATTORI DI RISCHIO, AREA EURO

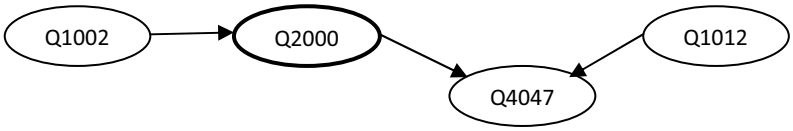


FIGURA 42:INSIEME NODI, ALGORITMO BNPC, DATASET FATTORI DI RISCHIO, AREA EURO

Sia la variabile che indica l'età dell'intervistato (Q1002), sia il tipo di risorsa utilizzata per cucinare (Q4047) sono in relazione di dipendenza diretta con la percezione dello stato di salute individuale come nel network riportato sopra. Si ha una seconda uguaglianza tra le due sotto-reti: la variabile che indica il lavoro corrente (Q1012) ha una dipendenza indiretta con Q2000.

Tra le due sotto-strutture apprese dai due differenti algoritmi esiste una notevole differenza: mentre nel primo caso le relazioni di dipendenza con i fattori di rischio sono molteplici ed indispensabili per individuare quali variabili giocano un ruolo fondamentale per la struttura del MB(Q2000), nel secondo caso questa aspetto non è presente; eliminando la dipendenza con la variabile Q4047, non sono presenti i fattori di rischio, ma solo variabili socio-demografiche.
Anche in questo caso le due strutture apprese portano a risultati diversi.

5.3.5 AREA GEOGRAFICA: SEARO

Il dataset che verrà qui analizzato è composto da 31315 unità statistiche che risiedono nel Sud Est-Asiatico.

Stato di salute

Di seguito sono riportate le analisi derivanti dall'insieme in cui si mettono in relazione le variabili socio-demografiche e quelle relative allo stato di salute.

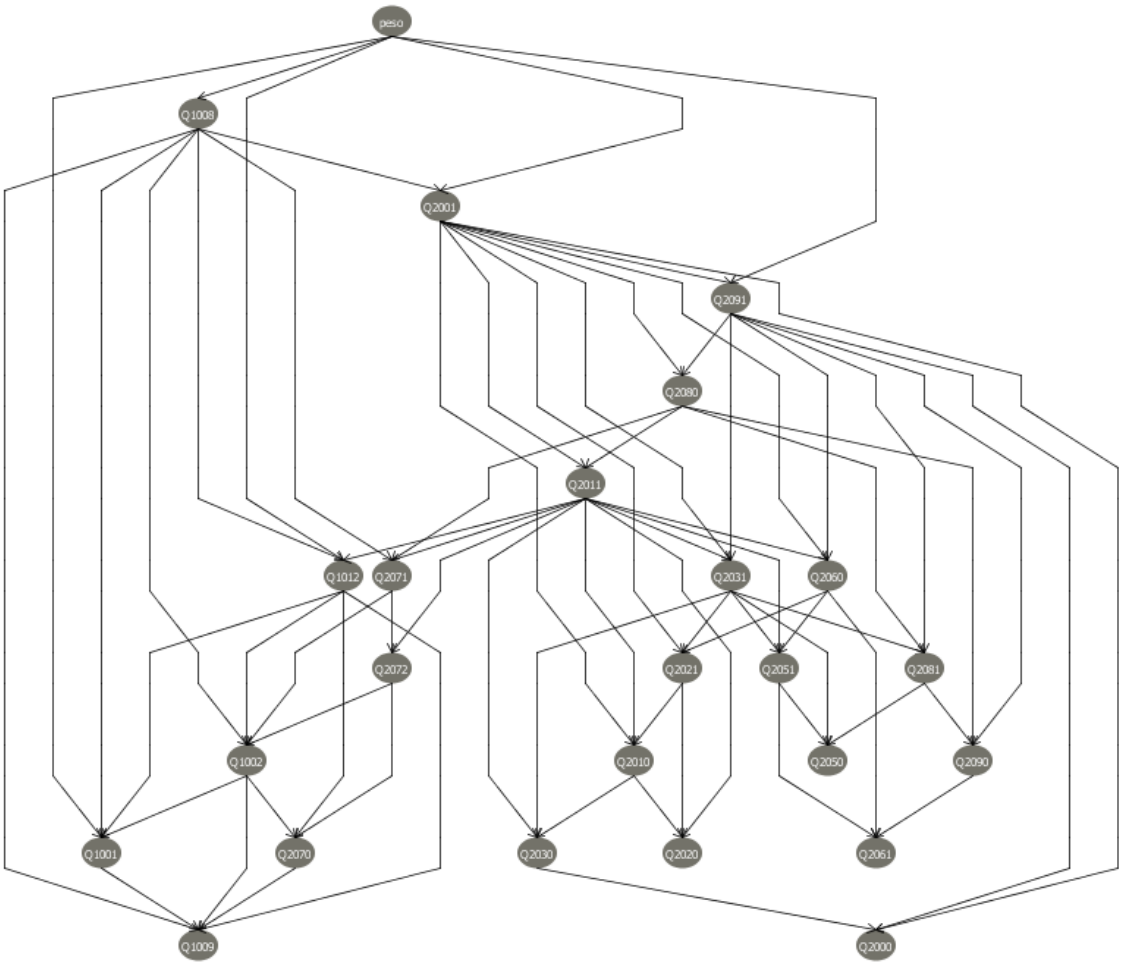


FIGURA 43:STRUTTURA APPRESA, ALGORITMO K2 , DATASET STATO DI SALUTE, AREA SEARO

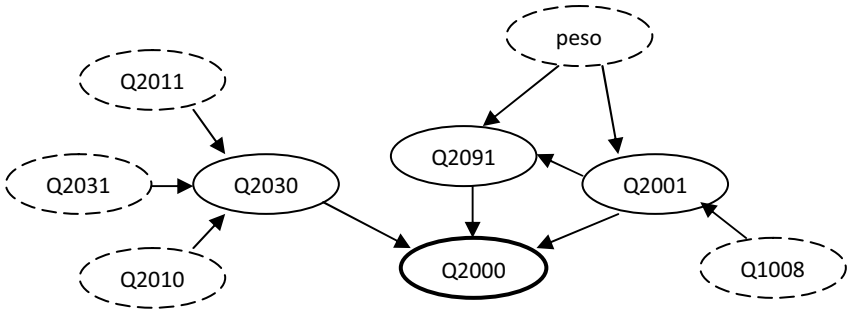


FIGURA 44:INSIEME NODI, ALGORITMO K2,DATASET STATO DI SALUTE, AREA SEARO

La percezione sullo stato di salute individuale è influenzato direttamente dalla difficoltà nello svolgere attività lavorative (Q2001), dalla quantità di dolori o fastidi fisici percepiti (Q2030) ed infine dalla sensazione di preoccupazione ed ansia. Nell'insieme riportato sono presenti le variabili Q2010 (difficoltà nei movimenti), Q2011 (difficoltà nello svolgere attività vigorose), Q2031(percezione di dolore fisico) e le due variabili socio-demografiche che descrivono lo stato matrimoniale (Q1008) ed il peso.

La struttura appresa mediante l'algoritmo BNPC, è invece il seguente:

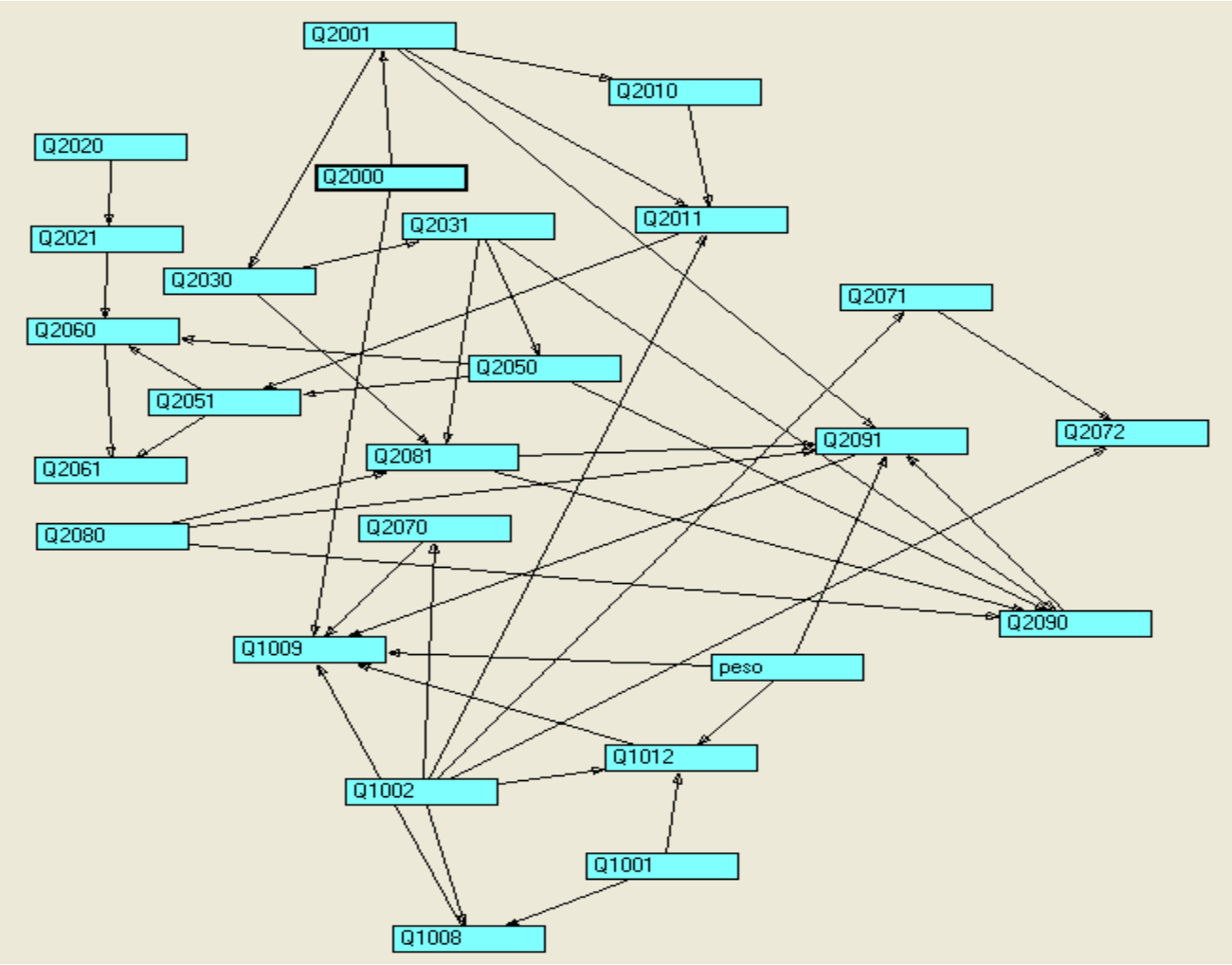


FIGURA 45:STRUTTURA APPRESA, ALGORITMO K2, DATASET STATO DI SALUTE, AREA SAERO

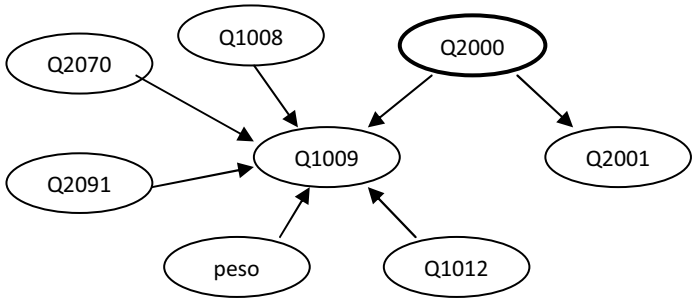


FIGURA 46:INSIEME NODI, ALGORITMO BNPC,DATASET STATO DI SALUTE, AREA SEARO

La struttura appresa dall'algoritmo BNPC è differente dal precedente: nonostante la variabile che indica la difficoltà nello svolgere attività lavorative (Q2001) sia in diretta relazione di dipendenza con Q2000, adesso essa svolge la funzione di figlio, mentre nel network precedente era un genitore del nodo. Viene identifica come variabile connessa alla percezione dello stato di salute il livello di educazione (Q1009), che a sua volta mette in relazione indiretta con Q2000 altre cinque variabili: il peso, lo stato matrimoniale (Q1008), il lavoro corrente Q(1012), la variabile che indica se un individuo indossa occhiali o lenti a contatto (Q2070) ed infine la sensazione di preoccupazione o ansia (Q2091).

Come si nota dai grafi e dalle descrizioni fatte, le due reti sono molto differenti: nel primo caso la percezione di salute ha principalmente collegamenti con le variabili relativi allo stato di salute, mentre la seconda rete considera anche le variabili socio-demografiche.

Fattori di rischio

In questo caso le osservazioni valide ammontano a 21708.

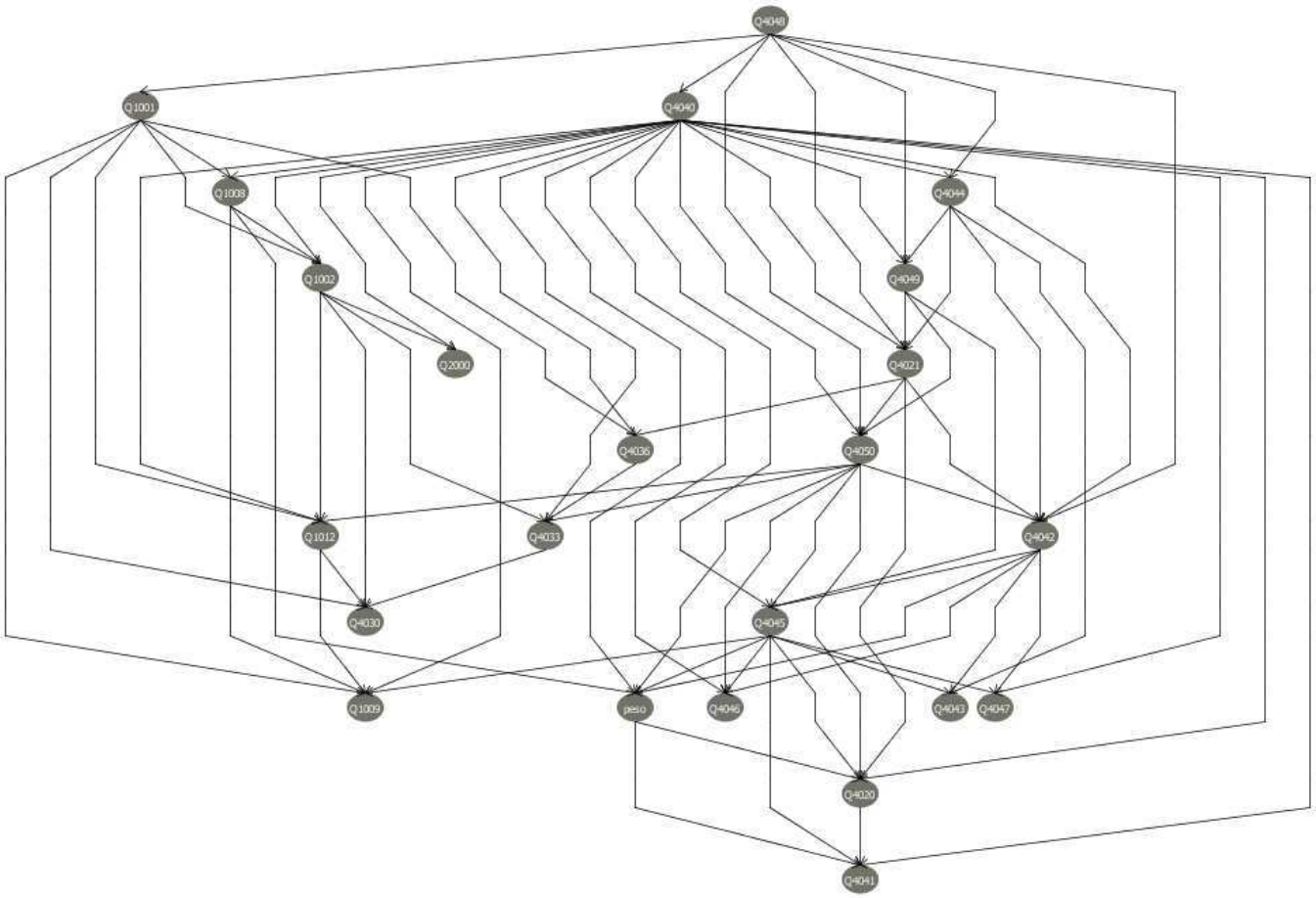


FIGURA 47:STRUTTURA APPRESA, ALGORITMO K2, DATASET FATTORI DI RISCHIO, AREA SEARO

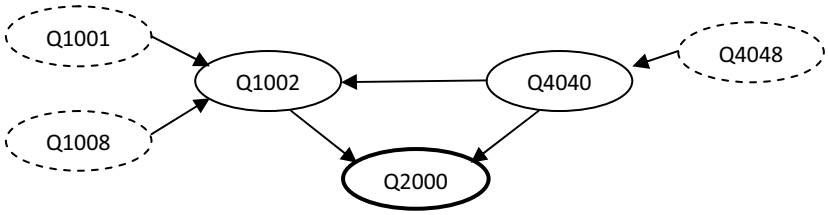


FIGURA 48:INSIEME NODI, ALGORITMO K2, DATASET FATTORI DI RISCHIO, AREA SEARO

Il markov blanket per Q2000 è formato da soli tre nodi, infatti la variabile d'interesse è in dipendenza diretta unicamente con l'età (Q1002) e con la variabile che indica il tipo di pavimenti presenti nell'abitazione (Q4040). Le altre variabili presenti non sono in connessione diretta e riguardano il sesso dell'individuo (Q1001), lo stato matrimoniale (Q1008) e il tipo di fornelli utilizzati nell'abitazione.
Sono quindi presenti sia variabili riferite ai fattori di rischio sia variabili socio-demografiche.

Il network bayesiano appreso dall'algoritmo BNPC è il seguente:

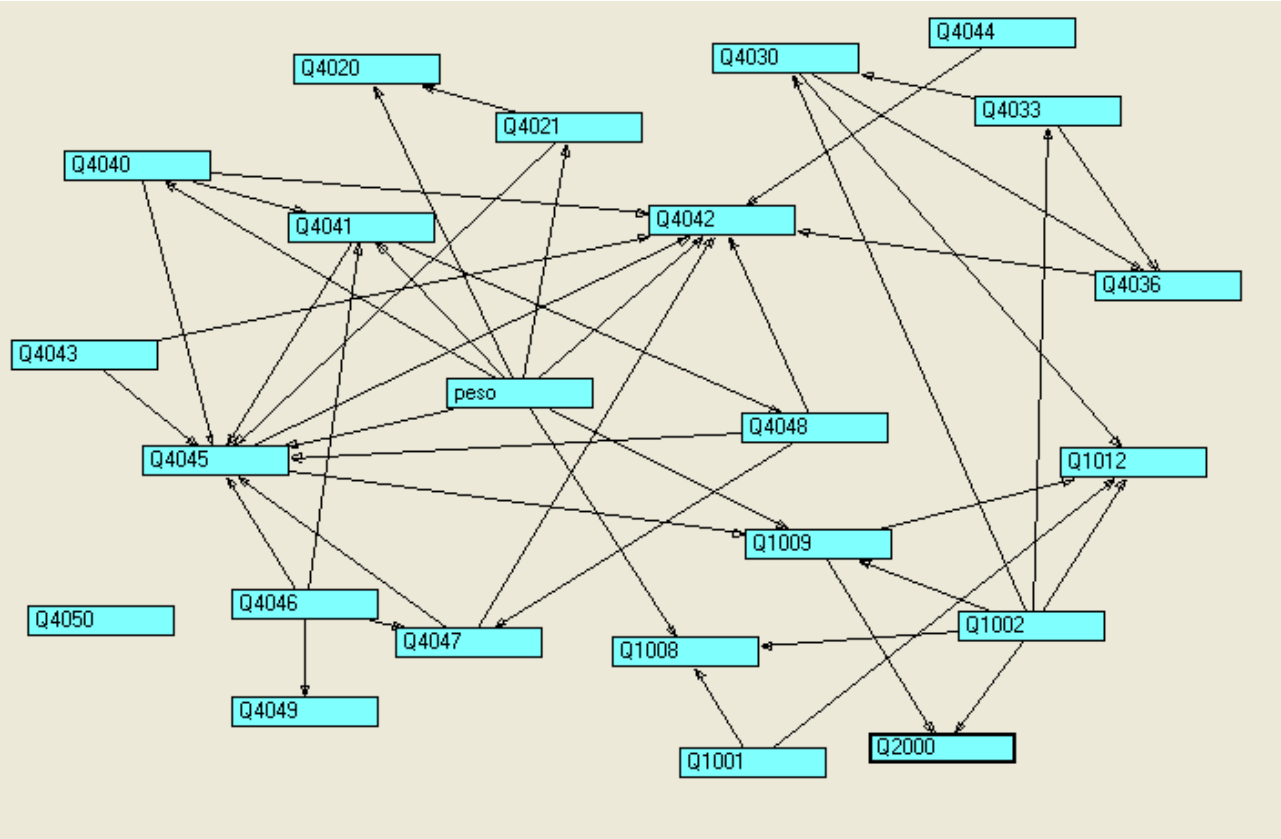


FIGURA 49:STRUTTURA APPRESA, ALGORITMO BNPC, DATASET FATTORI DI RISCHIO, AREA SEARO

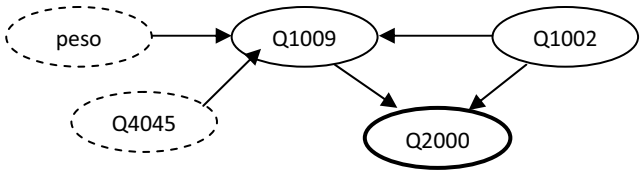


FIGURA 50:INSIEME NODI, ALGORITMO BNPC,DATASET FATTORI DI RISCHIO, AREA SEARO

La percezione dello stato di salute individuale dipende direttamente solo da variabili socio-demografiche; l'età (Q1002) ed il livello di educazione (Q1009) influenzano la variabile di interesse. Indirettamente possono essere considerate le variabili peso e il tipo di servizi igienici presenti nell'abitazione (Q4045).

Il network bayesiano precedente è differente da quello attuale per le variabili appartenenti all'insieme considerato; in entrambi però la Q2000 non presenta figli e sono presenti le variabili socio-demografiche. La differenza principale risiede nel MB(Q2000): nel primo è presente anche una variabile riferita ai fattori di rischio, nel secondo caso invece sono presenti solo variabili socio-demografiche.

5.3.6 AREA GEOGRAFICA: WPRO

Wpro è l'insieme dei paesi del Pacifico Occidentale composto da 16383 unità statistiche per la prima analisi, in cui si prendono in considerazione le variabili relative allo stato di salute.

Nella seconda parte, che prende in esame le variabili che descrivono i fattori di rischio, le osservazioni valide corrispondono a 8524.

Stato di salute

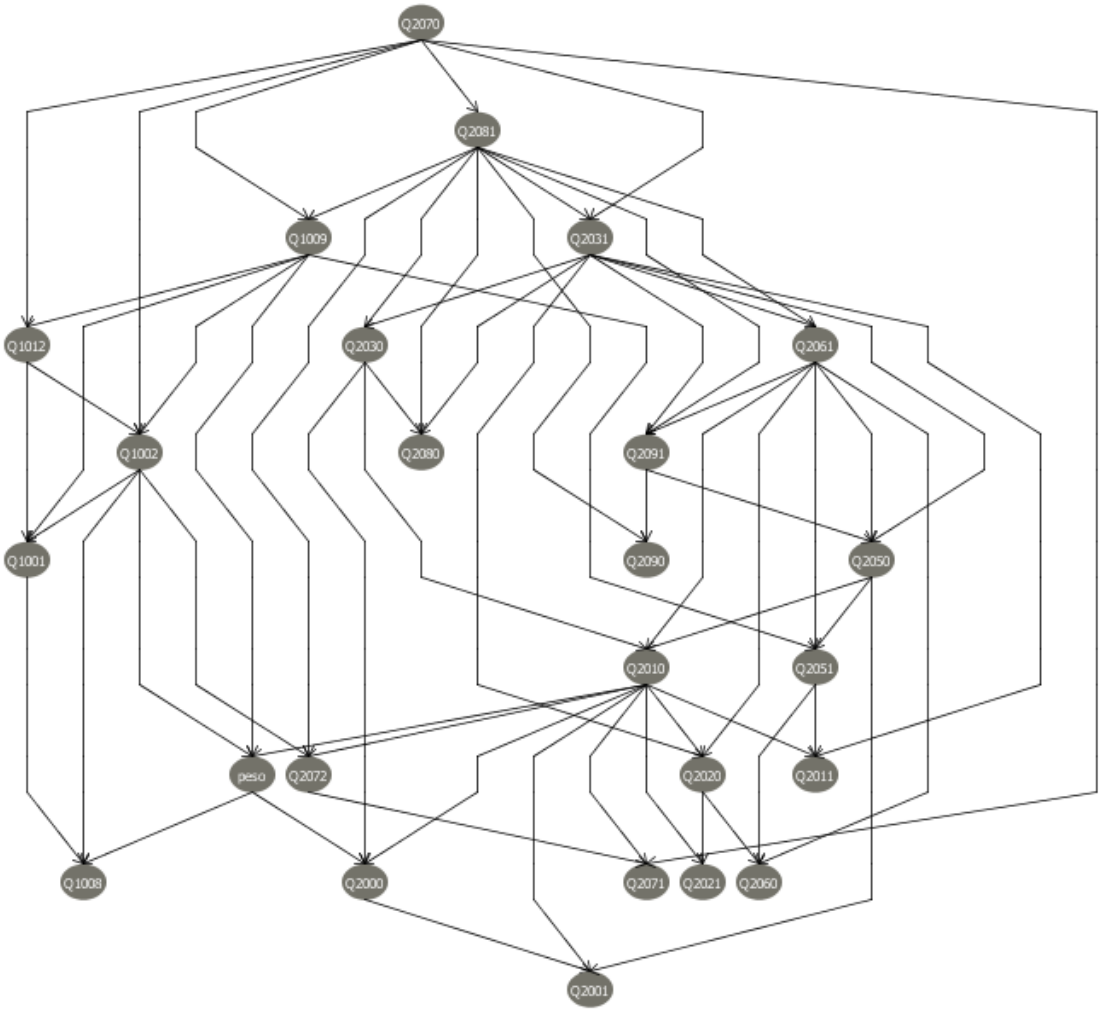


FIGURA 51: STRUTTURA APPRESA, ALGORITMO K2, DATASET STATO DI SALUTE, AREA WPRO

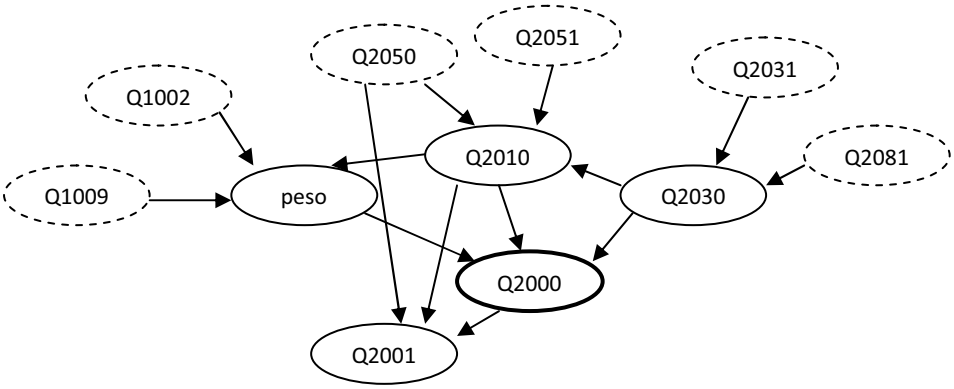


FIGURA 52:INSIEME NODI, ALGORITMO K2, DATASET STATO DI SALUTE, AREA WPRO

L'insieme costruito secondo le regole prima riportate produce una serie di dipendenze molto ampia: sono presenti tre variabili socio-demografiche di cui solo una, il peso, ha una relazione diretta con Q2000; le altre due riguardo l'età in anni e l'educazione dell'individuo intervistato. Tutte le rimanenti variabili appartengono invece all'insieme delle variabili riguardo lo stato di salute: quelle per cui esiste una relazione di dipendenza diretta riguardano la difficoltà nei movimenti (Q2010), la quantità di dolori o disturbi fisici (Q2030), le difficoltà nello svolgere attività lavorative (Q2001). Non appartenenti al Markov Blanket della variabile Q2000, ma riportate nell'insieme sono la difficoltà a ricordarsi o concentrarsi (Q2050), la difficoltà nell'imparare concetti nuovi (Q2051), la sensazione di dolori fisici (Q2031) ed infine la sensazione di stanchezza (Q2081). Si nota una struttura in cui hanno maggiormente peso le variabili che riguardano lo stato di salute, rispetto a quelle socio-demografiche.

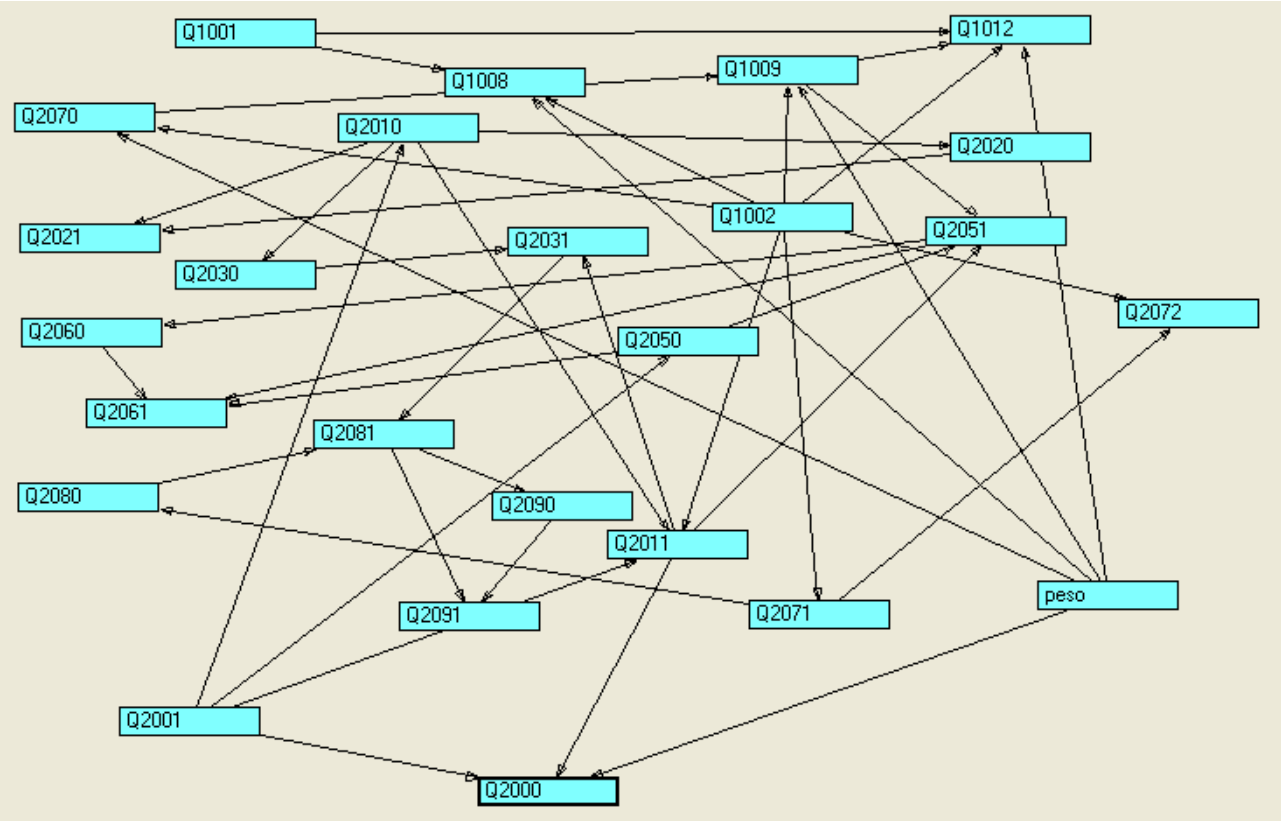


FIGURA 53:STRUTTURA APPRESA, ALGORITMO BNPC,DATASET STATO DI SALUTE, AREA WPRO

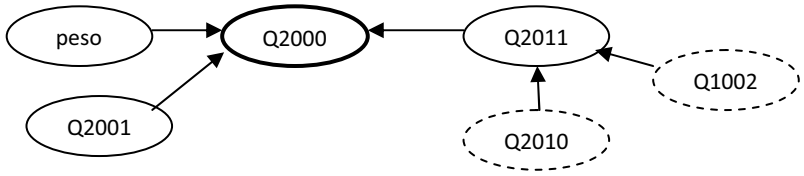


FIGURA 54:INSIEME NODI, ALGORITMO BNPC,DATASET STATO DI SALUTE, AREA WPRO

La struttura appresa dall'algoritmo BNPC risulta differente rispetto alla precedente. Il nodo d'interesse ha come genitori il peso, la difficoltà nello svolgere attività lavorative (Q2001) e la difficoltà nello svolgere attività vigorose (Q2011), che sono quindi in relazione diretta con la variabile d'interesse. Sono inoltre riportate le variabili riguardo l'età e riguardo le difficoltà nei movimenti che, al contrario, sono in relazione indiretta.

In entrambe le sotto-strutture la percezione dello stato di salute non presenta figli; in entrambi i casi sono presenti sia variabili sullo stato di salute sia relative allo stato socio-demografico dell'individuo.
La differenza però è evidente osservando il numero di relazioni che sussistono nelle due differenti situazioni; il numero delle variabile in esame è di gran lunga superiore nella prima parte dell'analisi.

FATTORI DI RISCHIO

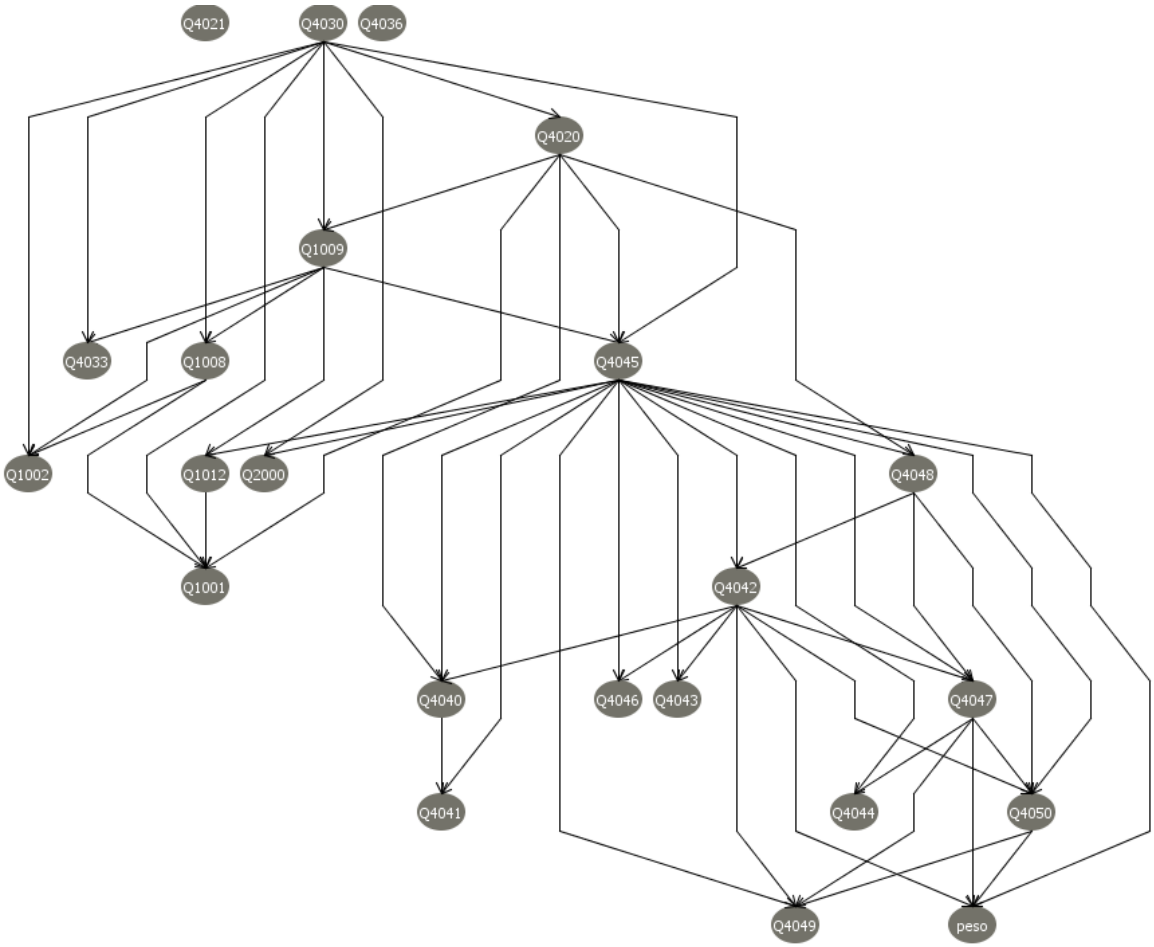


FIGURA 55:STRUTTURA APPRESA, ALGORITMO K2, DATASET FATTORI DI RISCHIO, AREA WPRO

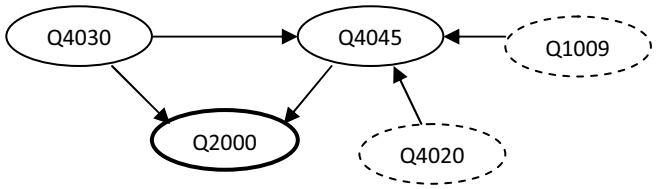


FIGURA 56:INSIEME NODI, ALGORITMO K2, DATASET FATTORI DI RISCHIO, AREA WPRO

La percezione dello stato di salute individuale dipende da quanti giorni di attività fisiche vigorose (Q4030) e dal tipo di servizi sanitari presenti nell'abitazione (Q4045). Tramite quest'ultima variabile risultano dipendenti direttamente anche il numero di frutti assunti giornalmente (Q4020) e il livello di educazione dell'individuo (Q1009).

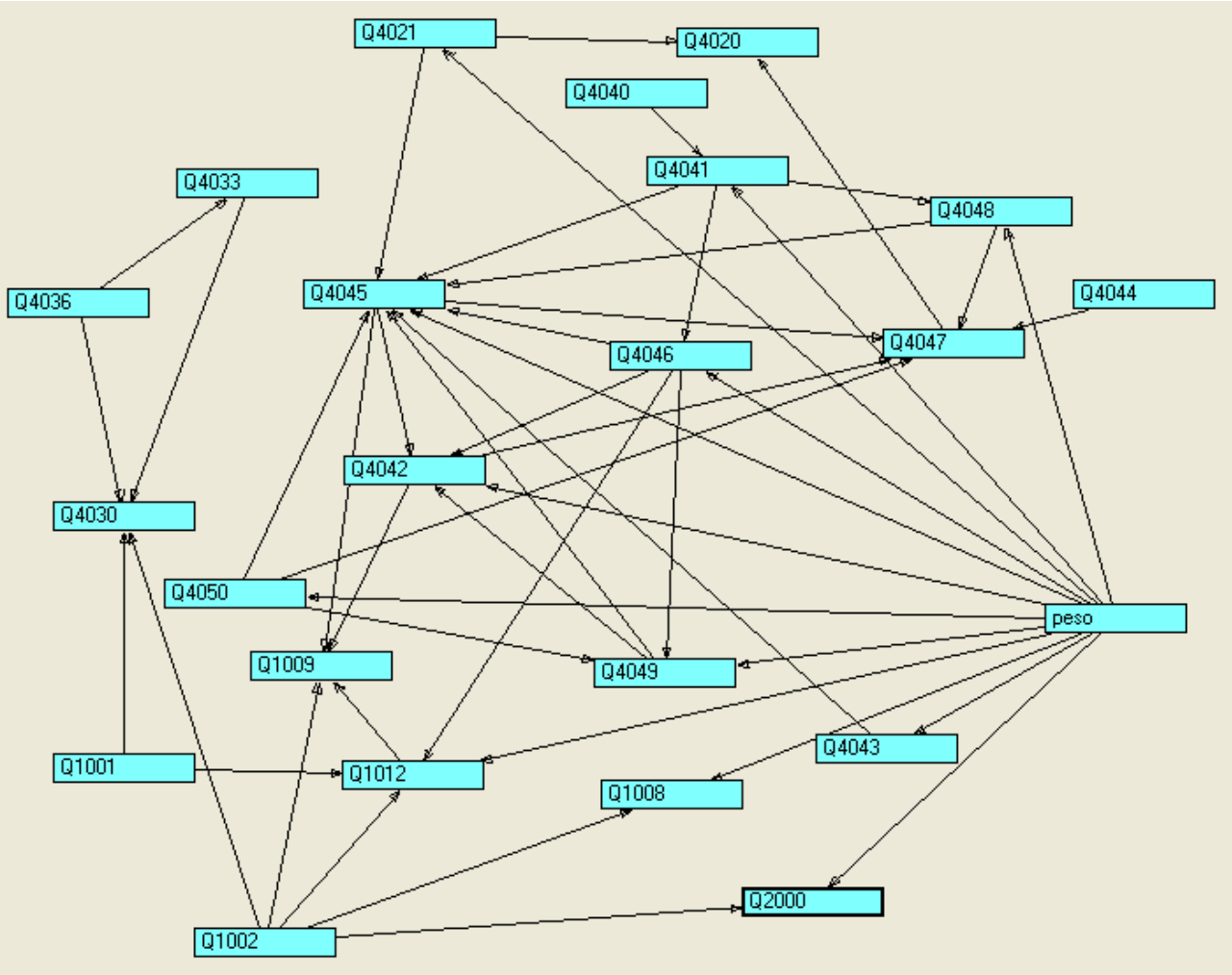


FIGURA 57:STRUTTURA APPRESA, ALGORITMO BNPC,DATASET FATTORI DI RISCHIO, AREA WPRO

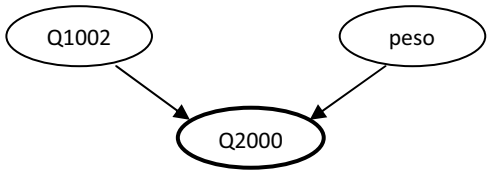


FIGURA 58:INSIEME NODI, ALGORITMO BNPC,DATASET FATTORI DI RISCHIO, AREA WPRO

La percezione della salute dipende direttamente dall'età (Q1002) e dal peso. Essi non sono considerati fattori di rischio, ma univocamente caratteristiche socio-demografiche.

Le due sotto-reti sono molto differenti tra loro: nel primo caso vengono prese in considerazione variabili riguardo lo stato di salute, mentre nel secondo caso sono presenti solo le variabili socio-demografiche.

CONCLUSIONI

Il network bayesiano è un modello grafico probabilistico che rappresenta una distribuzione di probabilità congiunta multivariata; è possibile dedurre da esse quali sono le in/dipendenze condizionate che sussistono tra le variabili di un dataset.

In queste analisi si è focalizzata l'attenzione sulla fase del learning della struttura di una rete bayesiana su un insieme di dati reali. L'obbiettivo della tesi è confrontare la struttura di una rete bayesiana appresa mediante l'utilizzo di due algoritmi appartenenti a due metodi di apprendimento differenti: l'approccio score-based seleziona la rete a cui è associato lo score più alto, confrontandolo secondo misure di bontà di adattamento. Come misura si è scelto di utilizzare lo score di Bayes.

I metodi constraint-based indagano sull'esistenza della in/dipendenza tra le variabili aleatorie presenti nel dataset; la dipendenza nel grafo viene tradotta con l'introduzione di un arco tra i due nodi in esame.

Lo svantaggio di quest'ultimo approccio riguarda la scarsa robustezza: piccole variazioni sui dati iniziali portano a rilevanti differenze nei risultati. Inoltre questi algoritmi sono esponenziali, ossia un numero elevato di variabili aleatorie implica un numero elevato di test da verificare.

Per perseguire l'obbiettivo si è proceduto con l'analisi di un dataset mediante l'ausilio del software Weka in cui è implementato l'algoritmo K2 ed utilizzando il software BNPC in cui è implementato l'omonimo algoritmo.

I dati utilizzati sono stati forniti dall'OMS (Organizzazione Mondiale per la Sanità): il dataset completo comprende 193057 osservazioni e quarantadue variabili relative alle caratteristiche socio-demografiche, lo stato di salute ed i fattori di rischio. La variabile d'interesse è la "percezione dello stato di salute individuale" (Q2000), i cui possibili stati sono cinque: "Very good", "Good", "Moderate", "Bad" ed infine "Very bed" . Il dataset originale è stato suddiviso in due sotto-campioni: uno per identificare quali variabili tra quelle socio-demografiche e quelle riguardo lo stato di salute sono in relazione di dipendenza con la variabile d'interesse, il secondo dataset con lo stesso obbiettivo è basato su un diverso insieme di variabili, costituito dall'insieme di variabili socio-demografiche e dalle variabili riguardo i fattori di rischio.

Dall'analisi, per ogni area geografica, si ricavano quattro network bayesiani:

1. Variabili riguardo lo stato di salute mediante l'utilizzo dell'algoritmo K2;
2. Variabili riguardo i fattori di rischio mediante l'utilizzo dell'algoritmo K2;
3. Variabili riguardo lo stato di salute mediante l'utilizzo dell'algoritmo BNPC;
4. Variabili riguardo i fattori di rischio mediante l'utilizzo dell'algoritmo BNPC.

Le strutture apprese dai due differenti algoritmi sono state confrontante tramite due insiemi ricavati nel seguente modo: si considera il markov blanket della variabile d'interesse ed i genitori dei genitori di Q2000. Mediante questa procedura è possibile mettere in evidenza le variabili che influenzano la percezione dello

stato di salute individuale indagando quali relazioni di dipendenza diretta o indiretta sussistono tra le variabili appartenenti all'insieme considerato.
In generale si può affermare che a partire da uno stesso dataset le strutture apprese dai due algoritmi risultano differenti. Se si considera unicamente il markov blanket, eliminando quindi i genitori dei genitori di Q2000, le relazioni di dipendenza apprese dai due algoritmi sono simili.
Si porta come esempio l'area geografica EMRO. Considerando il dataset composto dalle variabili riguardo lo stato di salute, sia dalla struttura derivante dall'esecuzione dell'algoritmo K2 che dell'algoritmo BNPC si evince che le variabili in relazione di dipendenza sono: la difficoltà nello svolgere attività lavorative (Q2001), la difficoltà nello svolgere attività vigorose (Q2011) e la difficoltà dei movimenti (Q2010).
Per alcune aree geografiche le uniche variabili per cui esistono delle relazioni di dipendenza con la variabile Q2000 sono esclusivamente quelle socio-demografiche; per altre aree geografiche, invece, si nota che le variabile di rischio o quelle riguardo lo stato di salute sono rilevanti. Ad esempio per i paesi appartenenti al continente Africano, le variabili per cui si può trovare una relazione con la percezione dello stato di salute individuale fanno principalmente parte dell'insieme dei fattori di rischio e dello stato di salute. La stessa situazione si ritrova nei paesi europei.
Per le aree geografiche EMRO e AMRO si nota invece che la percezione dello stato di salute dipende principalmente dalle variabili socio-demografiche. Al contrario per le aeree geografiche SEARO e WPRO questo non si riesce a definire in maniera così netta poiché sia le variabili socio-demografiche, i fattori di rischio e quelle relative lo stato di salute risultano importanti.

Tralasciando la suddivisione in aree geografiche si può affermare che l'età, la difficoltà nello svolgere attività lavorative, la difficoltà nel compiere movimenti, la difficoltà nel sostenere attività vigorose, il provare sensazioni di stanchezza, ansia o depressione sono le variabili che maggiormente influenzano la percezione dello stato di salute. Considerando i fattori di rischio, invece, le variabili che maggiormente influenzano la variabile d'interesse riguardano gli aspetti della vita quotidiana dell'individuo: il tipo di pavimento, di pareti, di servizi igienici presenti nell'abitazione.

APPENDICE

Di seguito sono riportate le definizioni dei termini richiamati nell'elaborato.

ATOMIC-COMPLETE:

Un insieme di distribuzioni P è atomic-complete per un insieme di grafi G se e solo se per ognumo dei grafi $g \in G$ e per ogni insieme disgiunto di vertici A, B, C esiste una distribuzione $p \in P$ tale che $A \perp B|C$ se e solo se $A \perp B|C$ è vero in $p \in P$.

MISURA DI LEBESGUE:

La misura di Lebesgue costituisce una generalizzazione del concetto elementare di volume dei sottoinsiemi dello spazio euclideo. È usata nella definizione dell'integrazione secondo Lebesgue. Gli insiemi a cui è possibile assegnare una misura di Lebesgue sono detti misurabili secondo Lebesgue; la misura dell'insieme Lebesgue-misurabile A è indicato con $\lambda(A)$. La costruzione moderna della misura di Lebesgue, basata sulle misure esterne, è dovuta a Carathéodory.

Per *ogni* sottoinsieme B di $\mathbf{R}^n$, possiamo definire

$$\lambda^*(B) = \inf\{\mathrm{vol}(M) : M \supseteq B, \text{e } M \text{ unione numerabile di prodotti di intervalli}\}.$$

Ora, vol(M) è la somma dei prodotti delle lunghezze degli intervalli coinvolti. Si definisce quindi l'insieme A misurabile secondo Lebesgue se

$$\lambda^*(B) = \lambda^*(A \cap B) + \lambda^*(B - A)$$

per tutti gli insiemi B. Questi insiemi Lebesgue-misurabili formano una σ-algebra, e la misura di Lebesgue è definita da $\lambda(A) = \lambda^*(A)$ per ogni insieme Lebesgue-misurabile A.

MAPPA PERFETTA:

Un grafo G è una mappa perfetta di una distribuzione P se

$$X \perp_P Y \mid Z \Leftrightarrow X \perp_G Y \mid Z.$$

COLLIDER:

Per ogni nodo adiacente ad un cammino, se due archi nel cammino nel loro punto finale incontrano in nodo v, si definisce v un nodo collider del percorso. Un nodo che non è un collider è denominato non-collider. Il concetto di collider si riferisce esclusivamente ad un solo cammino, per questo motivo, un nodo può essere un collider per un cammino particolare contemporaneamente un non-collider per un secondo percorso.

CROSS-VALIDATION:

La convalida incrociata consente di partizionare una struttura di un insieme di dati in sezioni trasversali, eseguire in maniera iterativa il training dei modelli e testarli a fronte di ciascuna sezione trasversale. È possibile specificare un numero di riduzioni in cui suddividere i dati; ciascuna riduzione viene quindi utilizzata come dati di test, mentre i dati rimanenti vengono usati per eseguire il training di un nuovo modello.

Si riportano le modalità di ogni variabile presente nel dataset analizzato:

- Variabili socio-demografiche (Respondent's Socio Demographic Characteristics):

  Paese di appartenenza (Country):

  AFRO, EURO, AMRO, EMRO, SEARO, WPRO

  Sesso (Q1001):

  M, F

  Età (Q1002):

  18-19, 20-29, 30-39, 40-49, 50-59, 60-69, 70;

  Stato matrimoniale (Q1008):

  N mar, C mar, S, D, W, , Co;

  Livello di educazione (Q1009):

  No, Less, Pri, Sec, Hi, Col, Pg;

  Lavoro corrente (Q1012):

  Ge, Nge, Se, E, Nw;

  Peso attribuito ad ogni unità statistica in base al disegno di campionamento (peso):

  m.bassa, bassa, media, alta, m.alta;

- Variabili sullo stato di salute (Health State Descriptions):

  Percezione stato di salute (Q2000):

  Very good, Good, Moderate, Bad, Very bad;

  Difficoltà nelle svolgimento di attività lavorative (Q2001):

  No, Mild, Moderate, Sever, Ex;

  Difficoltà nei movimenti (Q2010):

  No, Mild, Moderate, Sever, Ex;

  Difficoltà ad eseguire attività vigorose (Q2011):

  No, Mild, Moderate, Sever, Ex;

  Difficoltà nel vestirsi/lavarsi (Q2020):

  No, Mild, Moderate, Sever, Ex;

Difficoltà nel eseguire cure personali (Q2021):

No, Mild, Moderate, Sever, Ex;

Quantità dolori (Q2030):

No, Mild, Moderate, Sever, Ex;

Presenza dolori fisici (Q2031):

No, Mild, Moderate, Sever, Ex;

Difficoltà di concentrazione (Q2050):

No, Mild, Moderate, Sever, Ex;

Difficoltà ad imparare concetti nuovi (Q2051):

No, Mild, Moderate, Sever, Ex;

Difficoltà nelle relazioni interpersonali (Q2060):

No, Mild, Moderate, Sever, Ex;

Difficoltà nella gestione di conflitti e tensioni (Q2061):

No, Mild, Moderate, Sever, Ex;

Indossa occhiali o lenti a contatto (Q2070):

Yes, No;

Difficoltà nel vedere e riconoscere persone incontrate per strada (Q2071):

No, Mild, Moderate, Sever, Ex;

Difficoltà riconoscere oggetti (Q2072):

No, Mild, Moderate, Sever, Ex;

Problemi di insonnia (Q2080):

No, Mild, Moderate, Sever, Ex;

Percezione di sensazione di stanchezza (Q2081):

No, Mild, Moderate, Sever, Ex;

Sensazioni di tristezza o depressione (Q2090):

No, Mild, Moderate, Sever, Ex;

Sensazioni di ansia o preoccupazione (Q2091):

No, Mild, Moderate, Sever, Ex;

- Variabili riguardo a fattori di rischio (Risk Factors)

Quanti frutti al giorno consumati (Q4020):

Hard, Earth;

Quantità di verdura consumata al giorno (Q4030):

0,1,2,3,4,5,6,7;

Quanti giorni di attività fisica vigorosa/sportiva (Q4031):

0,1,2,3,…., 20;

Quanti giorni di attività fisica moderata (Q4033):

0,1,2,3,4,5,6,7;

Quanti giorni di attività fisica leggera/camminate (Q4036):

0,1,2,3,4,5,6,7;

Tipi di pavimento nell'abitazione (Q4040):

Hard, Earth;

Tipo di pareti nell'abitazione (Q4041):

C, M, T, P, Me,O;

Tipo di sorgente d'acqua potabile (Q4042):

Piped, Public, Tube, Pro Dug, Un Dug, Rain, Pond, Vendor;

Distanza dalla sorgente (Q4043):

Less, 5-30, , 30-60, 60-90, 90+;

Possibilità di avere 20 litri d'acqua potabile (Q4044):

Yes, No;

Tipo di sanitari presenti nell'abitazione (Q4045):

Piped, Septic, Pour, Cov Dry, Un Dry, Bucket, No, Other;

Distanza di sanitari dall'abitazione (Q4046):

Single, Multiple, Private, Shared;

Che tipo di risorsa viene utilizzata per cucinare (Q4047):

Gas, Eletr, Kerosene, Coal, Char, Wood, Crop, Animal, Grass, Other;

Tipo di fornelli (Q4048):

Without, With, Closed, Other;

Dove si cucina abitualmente (Q4049):

Used, Room, Separate, Outdoors, No;

Possibilità di riscaldare l'abitazione (Q4050):

Yes, No.

INDICE DELLE FIGURE

INDICE DELLE TABELLE

RINGRAZIAMENTI

Un ringraziamento va al CNR (National Research Council), istituto di neuroscienze di Padova, in particolare alla Professoressa Nadia Minicuci ed ad Alessandra Andreotti per la loro professionalità e disponibilità, per avermi dato l'opportunità di consultare e di analizzare dati appartenenti all'OMS (Organizzazione Mondiale della Sanità). Grazie alla stessa organizzazione per aver permesso di utilizzare questi dati.
Vorrei ringraziare tutti i professori che ho incontrato per aver condiviso le loro conoscenze e per avermi fatto capire quali sono i miei limiti e quali le mie doti.
Grazie alla Professoressa Adriana Brogini per avermi sempre incoraggiato ed aiutato nella stesura della tesi.

Grazie a tutti coloro con cui ho condiviso la mia esperienza padovana, per avermi fatto ridere, arrabbiare, pensare. In particolare grazie Ale per tutte le serate, i pomeriggi, le mattinate trascorse insieme tra una risata e la voglia di essere e sentirsi vivi.
Un ringraziamento va a Paolo per non aver ostacolato ed influenzato le mie scelte permettendomi di vivere serenamente questa esperienza. Grazie per essermi stato vicino in ogni momento.
Ma soprattutto un grazie speciale va i miei genitori e a mia sorella Chiara per avermi permesso di iniziare, proseguire e concludere questo importante cammino, ma soprattutto per aver sempre creduto in me e non aver dubitato mai, neanche un secondo, delle mie capacità.
Con tutto l'amore di cui sono capace, Federica.

BIBLIOGRAFIA

TESI CONSULTATE

D. Slanzi.
Reti bayesiane: Approcci per la selezione del modello.

M. Scutari.
Network Bayesiani: Un approccio non parametrico basato sull'entropia per la selezione del modello.

A. Andreotti.
Identificazione dei profili multidimensionali della salute: un'applicazione del modello grade of membership ai dati del world health.

D. Slanzi.
Bayesian belief network. Network bayesiano: metodologie e tecniche di analisi dell'incertezza.

T. Stich.
Bayesian network and structure learning (Observations on the sparse candidate algorithm and analyzation of questionnaires). University of Mannheim Department of mathematics and computer sciences. (2004)

D. Margaritis.
Learning Bayesian Network Model Structure from data. (2003)

ARTICOLI, DOCUMENTI E TESTI CONSULTATI:

R. Garey, S. Johnson.
Computers and Intractability: A Guide to the Theory of NP-Completeness. W.H. *Freeman (1979)*

R. Neapolitan.
Learning Bayesian network. (2004)

G. Cooper, E. Herskovits.
A bayesian method for the induction of probabilistic network from data. Machine learning 9, pag 309-347 (1992)

A. Azzalini.
Inferenza Statistica. Un'introduzione basata sul concetto di verosimiglianza. (1992)

L. Pace, A. Salvan.
Introduzione alla Statistica II. Inferenza, Verosimiglianza, Modelli. (2001)

A. Brogini.

Appunti del corso "Statistica Bayesiana".

A. Roverato.
Appunti del corso "Analisi dei dati categoriali".
J. Teugels, J. Horebeek.
Graphical Models for discrete data. Part 1: undirected graphs.

J. Teugels, J. Horebeek.
Graphical Models for discrete data. Part 2: acyclic directed graph.

A. Agresti.
Categorical Data Analysis. NY: Wiley. (1990)

D. E. Edwards.
Introduction to Graphical Modelling (2nd ed). Springer-Verlag, New York. (2000)

F.V. Jensen.
An Introduction to Bayesian Networks UCL Press, London. (1996)

R. Cowell, A. Dawid, S. Lauritzen, D. Spiegelhalter.
Probabilistic Networks and Expert Systems (1999)

S. Coles.
Appunti del corso "Statistica computazionale I".

S. Coles.
Appunti del corso "Statistica computazionale II".

R. Remco.
BouckaertBayesian Network Classifiers in Weka for Version 3-5-6.

D. Scuse, P. Reutemann.
WEKA Experimenter Tutorial for Version 3-5-7.

R. Kirkby, E. Frank, P. Reutemann.
WEKA Explorer User Guide for Version 3-5-7.

M. Hall, P. Reutemann.
WEKA KnowledgeFlow Tutorial for Version 3-5-7.

F. Eibe.
Machine Learning with WEKA.

B. Pfahringer.
Machine Learning with WEKA.

H. Akaike.
Information theory and an extension of the maximum likelihood principle.

S. Acid, l. de Campos, J. Fernàndez-Luna, S. Rodrìguez, J.M. Rodrìguez, J.L. Salcedo.

A comparison of learning algorithms for Bayesian networks: a case of study based on data from an emergency medical service.

X. Zhu, S. Zhang.
Methods for Generating Excel Files from SAS Datasets.The University of Mississippi, office of information technology and Mississippi center for supercomputing research. Importing MS Excel Data from SAS. (25 April 2007)

G. Della Vedova.
Imparare SAS. (21 Novembre 2003). Dipartimento di Statistica, Università di Milano-Bicocca.

M. Bolzan, A. Brogini, D. Slanzi.
Apprendimento di modelli grafici esplorativi per la valutazione in ambito socio sanitario: il caso dell'assistenza informale.

C. Meek.
Strong completeness and faithfulness in Bayesian networks.

A. Brogini, D. Slanzi.
On using BNs for complexity reduction in DTs.

L. M. De Campos.
Learning BN. Journal of machine learning research 7. (2006).

D. Slanzi.
A new approach for learning Bayesian Networks based on inference complexity.

J. Cheng, D. Bell, W. Liu.
Learning Bayesian Networks from Data: An Efficient Approach Based on Information Theory.

SITI CONSULTATI

http://ww.acm.org

http://auai.org

http://www.hugin.com

http://www.agena.co.uk

http://www.norsys.com

http://www.cs.berkeley.edu/~murphyk/Bayes/

http://www.cs.ualberta.ca/~jcheng/bnpc.htm

http://www.kddresearch.org/Groups/Probabilistic-Reasoning/

http://www.cs.Helsinki.FI/research/cosco

http://www.research.microsoft.com/research/dtg/bnformat/

http://www.cs.cmu.edu/~fgcozman/Research/Interchangeformat

http://www.cs.auc.dk/~marta/datamine.htm

http://www.cs.huij.ac.il/~galel

http://www.phil.cmu.edu/projects/tetrad/publications.html

http://bndev.sourceforge.net/

http://www.ics.uci.edu/~mlearn/MLRepository.html

http://leo.ugr.es/~elvira/

http://web.tiscali.it/mmariotti

http://sourceforge.net/project/stats/detail

http://www.wikipedia.it

http://www.cs.waikato.ac.nz/ml/weka/

http://wwwusers.ts.infn.it/~milotti/Didattica/Bayes/Bayes.html

http://www.math.unipd.it/~sperduti/SI08/apprendimento_bayesiano1.pdf

http://www-anw.cs.umass.edu/~cs691t/SS02/lectures/week7.PDF

http://www.doc.ic.ac.uk/~dfg/ProbabilisticInference/IDAPILecture11.pdf

http://citeseer.ist.psu.edu/darwiche95conditioning.html

http://www.cs.brown.edu/research/ai/dynamics/tutorial/Documents/GibbsSampling.html

http://www.cs.ubc.ca/spider/poole/papers/dynBN-UAI90.pdf

http://www.kddresearch.org/Workshops/RTDSDS-2002/papers/RTDSDS2002-GH-01.pdf

http://staff.icar.cnr.it/manco/Teaching/2006/datamining/lezioni/Lezione7.pdf

http://www.cs.ualberta.ca/~jcheng/bnpchlp/index.html

ftp://ftp.stat.umn.edu/pub/xlispstat/wip/

http://www.stat.umn.edu/ARCHIVES/archives.html

http://www.softpedia.com/dyn-search.php

http://www.brothersoft.com/downloads/bayes-estimator.html

http://www.brothersoft.com/downloads/bayes-estimator.html

http://www.kdnuggets.com/news/97/n28.html

http://www.stat.umn.edu/ARCHIVES/archives.html

http://www.cs.ualberta.ca/~jcheng/lab.htm

http://www.cs.ualberta.ca/~jcheng/resume.htm

http://weka.sourceforge.net/wekadoc/index.php/en:Explorer_-Classification_(3.4.6)

http://technet.microsoft.com/it-it/library/bb895194.aspx

http://technet.microsoft.com/it-it/library/bb895174.aspx

http://tesi.cab.unipd.it/archive/00004306/01/Scutari.pdf

http://en.wikipedia.org/wiki/Collider

http://www.cs.ualberta.ca/~jcheng/Doc/report98.pdf