PROGETTO DI UN SISTEMA INFORMATICO PER IL SUPPORTO ALLE DECISIONI MEDICHE

L'obiettivo del progetto è lo sviluppo un sistema informatico prototipale che supporti le attività di diagnosi, prognosi e terapia in ambito medico. Il sistema è basato sull'uso di metodologie e tecnologie di Basi di Dati, Machine Learning, Data Mining e Statistica Inferenziale.

1. Descrizione del contesto applicativo

I processi decisionali nella pratica clinica sono caratterizzati da una complessa articolazione che, schematicamente, può essere rappresentata in tre fasi:

- la formulazione della diagnosi dell'eventuale condizione patologica, effettuata attraverso l'osservazione dei sintomi ed identificando e rilevando opportuni dati biologici e clinici, a loro volta elaborati tramite la specifica conoscenza ed esperienza medica;
- la scelta e la pianificazione dell'intervento terapeutico migliore per un trattamento ottimale, in base a prefissati criteri, della malattia;
- l'analisi, la valutazione e la previsione dell'evoluzione della malattia nel tempo (prognosi), in modo da assumere gli eventuali interventi correttivi per quanto riguarda il trattamento terapeutico.

Ognuna di queste tre fasi è caratterizzata da una fondamentale funzione che consiste, essenzialmente, nell'identificare, acquisire, interpretare ed elaborare tutto un corredo di dati dal paziente. Tale funzione (comunque basata su protocolli e linee guida standardizzati, ma molto spesso svolta in modo soggettivo e qualitativo dal medico), proprio per la sua intrinseca natura, può essere resa più obiettiva, accurata ed affidabile attraverso l'ausilio di metodologie e tecnologie quantitative, opportunamente implementabili in sistemi computerizzati.

In generale, un Sistema di Supporto alle Decisioni Mediche (SSDM) deve, in senso opportuno, "istanziare" le conoscenze e l'esperienza del medico (l'esperto del dominio) e le sue capacità "problem-solving" nel risolvere i problemi decisionali riguardanti le varie attività cliniche. In tal senso, l'architettura del sistema è fondamentalmente basata su una "Base di Conoscenza Medica" e su un "Motore di Inferenza".

La Base di Conoscenza deve rappresentare, in modo opportuno, il corpo di conoscenze ed esperienze dello specifico dominio; a tal fine, tramite una stretta collaborazione con gli esperti del dominio, occorre sviluppare procedure per l'acquisizione e l'opportuna codifica della conoscenza medica di interesse.

Il Motore d'Inferenza deve, invece, implementare, attraverso adeguate metodologie, il mix di approcci ipotetico – deduttivo ed induttivo che è alla base del processo di inferenza della decisione medica. Dal punto di vista strettamente metodologico, nella stragrande maggioranza delle applicazioni, il motore di inferenza implementa il cosiddetto problema della Classificazione, nel quale occorre decidere circa l'appartenenza di un soggetto ad uno fra i diversi insiemi di una famiglia. A tal fine, ogni soggetto è caratterizzato da un prefissato insieme di attributi, i cui specifici valori determinano l'appartenenza ad una classe. Ad esempio, nel caso di un problema di Diagnosi, la decisione medica consiste nel discriminare, relativamente alla data patologia, se il paziente sotto esame si trova in uno stato di malattia o meno, quindi se appartiene alla classe "pazienti sani" o alla classe "pazienti malati". Il processo di classificazione si realizza, pertanto, determinando un "criterio di separazione" delle due classi di pazienti, criterio che viene definito "addestrando" il classificatore (ovvero il Motore d'Inferenza) sulla Base di Conoscenza, che in questo caso viene costruita considerando un opportuno campione di pazienti sani e malati, ognuno dei quali è caratterizzato da un ben definito insieme di attributi, che identificano i segni, i sintomi, i valori di parametri clinici e biologici specifici della data patologia. Il "criterio di separazione" così determinato, sarà, quindi, utilizzato per discriminare un nuovo caso che si presenterà all'attenzione del medico.

1.1 **Diagnosi precoce del Tumore al Seno**

La valutazione delle caratteristiche morfologiche dei nuclei cellulari provenienti da campioni d'aspirato con ago sottile nonché la diagnosi dipende esclusivamente dall'esperienza clinica dello specialista, il quale si avvale di un approccio di tipo ***associativo*** (cioè del confronto tra il nuovo caso ed altri già noti) per stabilire se il campione di "liquido" è stato prelevato da un nodulo benigno o maligno. Risulta molto utile avere a disposizione un Sistema Diagnostico Automatico

Computerizzato (CAD) che sia d'ausilio allo specialista nelle fasi di rilevamento, esame ed interpretazione dei dati citomorfologici provenienti da campioni d'aspirato con ago sottile.
Allo scopo di arrivare allo sviluppo ed implementazione di un sistema di questo tipo, bisogna partire dalla procedura diagnostica basata ***sull'esame citologico per ago aspirato.***
Dall'esame di questo diagramma risulta evidente la possibilità di costruire un ***sistema computerizzato di supporto alla decisione diagnostica*** mediante l'integrazione di due moduli:
Un modulo per l'elaborazione interattiva d'immagini digitalizzate riprese al microscopio;

- Un modulo per la discriminazione ("classificazione") automatica dei campioni d'aspirato con ago sottile in due classi, quella benigna e quella maligna.

Il primo modulo deve implementare le seguenti funzioni:
Acquisizione d'immagini citologiche (relative ai vetrini opportunamente scelti e colorati) digitalizzate attraverso una configurazione microscopio-telecamera;

- Rilevazione e misurazione delle caratteristiche citologiche dei nuclei cellulari selezionati.

Il secondo modulo deve "implementare l'esperienza clinica", ovvero deve essere capace di "replicare", in maniera automatica, il processo di diagnosi realizzato dal medico specialista sulla base della propria esperienza clinica. Osservando che il processo di diagnosi può essere interpretato come un processo decisionale basato sull'esame d'opportuni dati biologici e clinici, è evidente che, in tale processo, assume un ruolo decisivo la fase di "classificazione", laddove per *classificazione* s'intende la *funzione di discriminare l'eventuale stato patologico tramite l'elaborazione e l'interpretazione di quegli attributi che, in qualche misura, descrivono lo stato della patologia sotto esame.*
L'acquisizione dei dati comporta i seguenti passi:
1**.** Scelta della colorazione dei vetrini.
2. Ricerca e selezione dei vetrini continenti i nuclei delle cellule campione.
3. Acquisizione delle immagini digitalizzate attraverso una configurazione microscopio-telecamera.
4. Selezione dei nuclei campione.
5. Rilevazione e misurazione delle caratteristiche citologiche dei nuclei.
Le caratteristiche morfologiche da identificare e misurare sono da riferirsi alla dimensione,alla forma ed ai livelli di cromatina del nucleo.
I moduli di misurazione utilizzati sono stati:
1**.** Misurazione densità ottica
2. Misurazione automatica multicanale

<u>Misurazione densità ottica</u>
Tale modulo dì misurazione permette di effettuare misurazioni sulla assorbanza e trasmittanza di un corpo analizzando immagini in scala di grigi. Nella scala di grigi i valori sono compresi tra 0 e 255 cui corrisponde rispettivamente il bianco e il nero.
I parametri densitometrici rilevati nell'oggetto selezionato sono:

- AREA (A) = numero totale di pixel
- GRIGIO MEDIO (G.ME) = media dei livelli di grigio dei pixel
- GRIGIO MINIMO (G.MI) = livello minimo di grigio tra i pixel
- GRIGIO MASSIMO (G.MA) = livello massimo di grigio tra i pixel
- GRIGIO INTEGRATO (G.INT) = somma dei livelli di grigio dei pixel
- DEVIAZIONE STANDARD (DEV.ST) = deviazione standard dei livelli di grigio dei pixel
- DENSITA' OTTICA (D) = rapporto tra grigio integrato ed area: D = G.INT / A

<u>Misurazione automatica multicanale</u>
Tale modulo di misurazione permette di effettuare misurazioni di carattere geometrico su un corpo analizzando immagini a colori**.**
I parametri geometrici rilevati nell'oggetto selezionato sono:

- AREA (A) = numero totale di pixel

DIAMETRO DEL CERCHIO EQUIVALENTE (DCE) = diametro del cerchio avente la stessa area dell'oggetto

- DIAMETRO DI FERET = dimensione massima in una direzione specifica
- LARGHEZZA (LA) = larghezza del diametro di Feret più corto (minimo fra i diametri calcolati a 0°, 12.5°, 35°, 57.5°, 22.5°, 45°, 67.5°, 90°)
- LARGHEZZA FIBRA (LAF) = lunghezza del lato più corto del rettangolo avente la stessa area e perimetro dell'oggetto misurato (è un'approssimazione della larghezza di fibre curve): LAF = [P -] /4
- LUNGHEZZA (LU) = lunghezza del diametro di Feret più lungo (massimo fra i diametri calcolati a 0°, 12.5°, 35°, 57.5°, 22.5°, 45°, 67.5°, 90°)
- LUNGHEZZA FIBRA (LUF) = lunghezza del lato più lungo del rettangolo con la stessa area e perimetro dell'oggetto misurato (approssima la lunghezza di fibre curve):

LUF = [P +] /4

- PERIMETRO (P) = somma dei pixel appartenenti alla linea che circoscrive l'oggetto
- RAPPORTO D'ASPETTO (RA) = rapporto tra lunghezza ed ampiezza: RA = LU / LA
- ROTONDITA' (R): R = P2 /(4*л*A*1.064

Descrizione dei dati

Sono state acquisite 90 immagini da microscopio di cui 31 relative a nuclei cellulari benigni e 59 relative a nuclei cellulari maligni. Da queste immagini si sono ricavati e misurati 550 nuclei benigni e 575 nuclei maligni, opportunamente selezionati in modo da coprire, omogeneamente, tutto "l'universo" della patologia in esame.

Si è così generata una base di dati costituita da 1125 casi ciascuno dei quali rappresentato da una tupla (vettore) con 13 attributi (componenti). Ogni attributo identifica uno dei caratteri citomorfologici che il medico specialista può considerare al fine di stilare la sua diagnosi, e che si riferiscono alla dimensione, alla forma ed ai livelli di cromatina nel nucleo. Nel caso in esame, per ogni nucleo cellulare, sono stati selezionati 7 attributi geometrici (provenienti dalla misurazione multicanale) e 6 attributi densitometrici (provenienti dalla misurazione della densità ottica).

Gli attributi geometrici sono:

- *Area (A)* del nucleo
- *Diametro del cerchio equivalente* (*DEq*) avente la stessa area del nucleo
- *Lunghezza (Lung)* del diametro di Feret più lungo
- *Larghezza (Larg)* del rettangolo equivalente con stessa area e stesso perimetro del nucleo
- *Perimetro (P)* del nucleo
- *Rapporto d'aspetto (RA)*
- *Rotondità (R)*

Gli attributi densitometrici sono:

- *Grigio Minimo (GMin)*
- *Grigio Medio (GMed)*
- *Grigio Max (GMax)*
- *Grigio Integrato (GInt)*
- *Deviazione Standard (DevSt)*
- *Densità Ottica (Dens)*

Il primo gruppo di attributi consente di valutare le alterazioni dì dimensione e forma del nucleo; il secondo gruppo, invece, consente di valutare le alterazioni della cromatina.

1.2 Diagnosi precoce di Infarto Miocardio Acuto

L’obiettivo di tale diagnosi è la distinzione tra pazienti affetti da dolore toracico e pazienti che sono stati colpiti da infarto miocardio acuto o che sono a rischio di infarto.
Il data set utilizzato per gli esperimenti computazionali riguarda pazienti affetti da malattie cardiache, e contiene casi relativi a pazienti soggetti a Infarto Miocardio Acuto (IMA) o dolore toracico puro. Sono presenti 242 casi, contraddistinti da 105 caratteristiche
Le caratteristiche del paziente sono state divise in 4 categorie.
La prima categoria è quella dei “dati anamnestici”, che comprende:

- Sesso
- Età
- Familiarità
- Sudorazione
- Dispnea
- Pressione massima e minima
- Fumo
- Ipertensione
- Peso
- Diabete
- Dislipidemia

Nella seconda categoria, denominata “elettrocardiogramma”, vi sono valori quali:

- Ritmo sinusale
- Frequenza cardiaca
- Andamento ST superiore ed inferiore
- Andamento onda T negativa
- Andamento onda Q
- PQ
- QRS
- QTC
- Ipertrofia

Vi sono poi i dati relativi alle analisi del sangue, cioè gli “esami ematochimici”:

- Creatina
- GOT
- GPT
- LDH
- CPKMB
- Troponina
- Proteina C reattiva
- Glicemia
- Fibrinogeno
- Ves
- Colesterolo
- Triglicedi
- Uricemia
- Azotemia
- Sodio
- Potassio
- Emoglobina
- Ematocrito
- Globuli rossi

Nell’ultima denominata “ecocardiogramma”, vi sono dati relativi a:

- Ipocinesia
- Acinesia
- Discinesia
- Indice cardiotoracico
- Diametri SIV, PP, DD, DS
- Pericardio

2. Servizi offerti dal sistema informatico

Gli utenti (**operatori medici**) possono accedere al sistema attraverso un'opportuna interfaccia web, dopo essere stati riconosciuti mediante l'inserimento di codice identificativo e password.

Un operatore può **ottenere informazioni riguardo all'organizzazione**: in particolare, egli può visionare l'elenco dei vari operatori medici registrati sul sistema informatico, con indicazione dei rispettivi ruoli e reparti di appartenenza (Oncologia, Cardiologia) .

Un operatore può effettuare semplici analisi statistiche sfruttando un insieme **di interrogazioni predefinite sui dati storici relativi a casi medici**.

Le principali interrogazioni da effettuare (sui dati relativi all'infarto) sono le seguenti:

- numero di pazienti malati selezionati in base ad un range di età (e/o di peso), scegliendo tramite form i valori minimo e massimo per l'età (e/o il peso);
- numero di pazienti malati, raggruppati in base ad uno o più attributi anamnestici di tipo categorico (sesso, familiarità, fumo, …)
- valori medio, minimo e massimo di una misura (attributo) di esami ematochimici per un insieme di pazienti selezionati in base a range di età e/o peso e raggruppati in base ad uno o più attributi anamnestici di tipo categorico (sesso, familiarità, fumo, …)

Il sistema offre, inoltre, agli operatori un insieme di servizi per creare e usare nuova conoscenza, in forma di **modelli di supporto alla diagnosi,** cioè modelli di classificazione che consentono di prevedere la presenza/assenza di malattia per un nuovo caso. Tali modelli di classificazione possono essere costruiti in modo automatico, applicando un algoritmo di data mining (per l'induzione di classificatori) su un dataset di esempi, cioè un insieme di casi già di cui è nota la classe usati per l'addestramento (training) del modello.

Il sistema deve, pertanto, consentire all'operatore di **costruire un nuovo dataset** derivato dai dati medici originari, selezionando il dominio di dati (*Miocardio* o *TumoreSeno*) e gli attributi da considerare.

Un operatore può **generare un nuovo modello** in modo automatico selezionando un dataset, un algoritmo di mining e impostando opportunamente i parametri dell'algoritmo. La data di creazione del modello deve essere impostata automaticamente.

Successivamente, l'operatore può visualizzare l'elenco dei modelli da lui creati ed effettuare una qualunque delle azioni seguenti:

- **visualizzare dati dettagliati su un modello** (stime di accuratezza, matrice di confusione, eventuale rappresentazione grafica/testuale del modello, informazioni sul dataset utilizzato per l'addestramento del modello, valori dei parametri utilizzati nell'addestramento, …)
- **eliminare un modello**
- **usare un modello per effettuare una predizione**

Un operatore può infine **memorizzare una diagnosi** per un caso, facendo eventualmente riferimento alla predizione fornita per tale caso da uno dei modello di classificazione a sua disposizione.

2.1. Generazione di un dataset

Per generare un nuovo modello di classificazione è necessario fornire creare un dataset costituito da casi di cui è nota la classe di appartenenza (*training set*).

A tale scopo, l'operatore deve indicare:

- la sorgente dati (tabella *Miocardio* o tabella *TumoreSeno*) da cui estrarre il dataset
- il numero di casi (tuple) da estrarre dalla sorgente
- il tipo di selezione degli attributi: automatica o manuale
- il nome del file (in formato "arff", specificato nel seguito) in cui memorizzare il dataset

Successivamente, l'utente deve selezionare un sottoinsieme degli attributi della sorgente scelta:

- nel caso di selezione manuale, l'utente può scegliere direttamente gli attributi da un elenco visualizzato dall'interfaccia;
- nel caso di selezione automatica, invece, l'utente deve fornire i parametri richiesti dall'algoritmo di feature selection impiegato.

Il **formato ARFF** è usato dalla libreria WEKA per rappresentare un dataset mediante un file di testo. Di seguito è mostrato un esempio di file ARFF che descrive un dataset di 14 casi: ogni caso rappresenta una giornata ed è caratterizzato da attributi relativi alle condizioni climatiche e dall'attributo di classe *play* (con due valori possibili *yes* e *no*, corrispondenti, rispettivamente, alla classe delle giornate in cui si può giocare e a quella delle giornate in cui non si può giocare):

```
@relation weather

@attribute outlook {sunny, overcast, rainy}
@attribute temperature real
@attribute humidity real
@attribute windy {TRUE, FALSE}
@attribute play {yes, no}

@data
sunny,85,85,FALSE,no
sunny,80,90,TRUE,no
overcast,83,86,FALSE,yes
rainy,70,96,FALSE,yes
rainy,68,80,FALSE,yes
rainy,65,70,TRUE,no
overcast,64,65,TRUE,yes
sunny,72,95,FALSE,no
sunny,69,70,FALSE,yes
rainy,75,80,FALSE,yes
sunny,75,70,TRUE,yes
overcast,72,90,TRUE,yes
overcast,81,75,FALSE,yes
rainy,71,91,TRUE,no
```

La prima riga indica il nome del dataset preceduto dal tag @relation. Seguono le descrizioni dei vari attributi, riportate una per riga e precedute @attribute: per ogni attributo è specificato il nome ed il tipo, che può essere numerico (*real*) o categorico. Nel caso di attributi categorici è necessario specificare i valori che l'attributo può assumere, fra parentesi graffe e separati da virgole.
Per default si assume che l'ultimo attributo rappresenti la classe e, pertanto, deve essere di tipo categorico.
Il tag @data introduce, infine, l'elenco dei casi del dataset: ogni caso (tupla) corrisponde ad una riga contenente i valori degli attributi del caso, separati da virgole. L'ordine dei valori corrisponde all'ordine secondo il quale sono stati definiti precedentemente gli attributi.

3. Implementazione

Il sistema informatico può essere realizzato secondo un'architettura distribuita, comprendente le seguenti tipologie di componenti:

- una base di dati
- un interfaccia web per gli utenti
- moduli per l'implementazione della logica applicativa (accesso ai servizi informativi e di analisi)

La base di dati ha lo scopo di permettere la memorizzazione, l'interrogazione e la manipolazione dei dati e delle altre informazioni utili per l'erogazione dei servizi agli utenti e per la gestione del sistema.
L'interfaccia consiste in un sito web e può essere realizzata definendo un insieme di pagine statiche e dinamiche (JSP).
L'implementazione delle varie operazioni di gestione e di accesso alle informazioni e ai servizi del sistema può avvantaggiarsi dell'impiego di opportuni moduli (classi Java), richiamabili nelle componenti di interfaccia. In particolare, si consiglia di sfruttare i servizi offerti dalla libreria JDBC, per l'accesso a database, e dalla libreria WEKA, per operazioni di data mining (generazione di modelli di classificazione e predizione su nuovi casi).