UNIVERSITÀ DI PISA

Corso di Laurea in Informatica Umanistica

RELAZIONE

SUPERVISED LINK PREDICTION SU RETI SOCIALI

Candidato: *Federico Martinelli*

Relatore: *Dino Pedreschi*

Correlatore: *Giulio Rossetti*

Anno Accademico 2013-2014

"A mio padre, semplice e profondo, che sicuramente vivrà questo momento con il suo sorriso sornione lanciando applausi dal paradiso.

A mia madre, il bianco e l'indaco del cielo e l'acqua che mi disseta, da cui ho preso il cuore e non l'ha voluto indietro.

A mia zia, il suo sorriso e e la sua leggerezza, che mi ha fatto capire che l'allegria va consumata anche per un solo momento. "

Indice

INTRODUZIONE

Negli ultimi anni la comunità scientifica ha manifestato un grande interesse per l'analisi e l'estrazione di conoscenza da sistemi complessi.
Spesso realtà complesse, anche molto diverse tra loro, possono essere efficacemente descritte in termini di reti: insiemi di entità connesse tra di loro attraverso una qualche relazione. Esempi ne sono le reti sociali, biologiche, tecnologiche ed economiche. In tutti questi casi, la caratteristica comune è l'esistenza di proprietà d'interconnessione che danno luogo a complesse topologie.

In questo lavoro di tesi concentreremo la nostra analisi su di uno specifico contesto: analizzeremo solamente *"reti sociali"*. Tali oggetti modellano l'interazione tra le persone appartenenti ad un gruppo o comunità: in questo scenario, ciascun nodo rappresenta una persona mentre gli archi, che connettono coppie di nodi, una qualche forma di associazione tra persone.
Le reti sociali sono oggetti molto dinamici, poiché nuovi archi e nodi possono essere aggiunti (ed anche rimossi) al grafo nel tempo. Comprendere le dinamiche che guidano l'evoluzione delle reti è il primo passo per ottenere intuizioni sulla vera natura del fenomeno di volta in volta osservato.

Il problema che andremo ad analizzare, traendo spunto dall'articolo di Liben-Nowell e di Kleinberg *"The Link Prediction Problem for Social Network"*, è quello di prevedere le nuove connessioni che si instaureranno tra i nodi di una rete. Questo problema è chiamato Link Prediction. L'obiettivo ultimo di approcci formulati per il Link Prediction è quello di prevedere come evolverà la struttura della rete analizzata sfruttando le informazioni topologiche da essa fornite.

Tale problema viene usualmente affrontato tramite due tipologie di approcci: può essere infatti risolto sia tramite una analisi *non supervisionata* che *supervisionata*.
Nella prima tipologia d'indagine rientrano tutti quei metodi che, per

predire la presenza di nuovi archi, assegnano uno score di confidenza a ciascuna coppia di nodi usando come input informazioni estratte dalla rete; i risultati poi vengono ordinati in una classifica, ordinata secondo un punteggio decrescente, in modo da poter evidenziare gli archi la cui predizione è ritenuta più affidabile. In questo approccio si può notare come le informazioni sulla topologia della rete siano estratte in modo diretto e mirato soggetto ad una scelta esplicita dell'utente. L'utente decide a priori quale caratteristica della rete si presta meglio a spiegarne l'evoluzione: la bontà delle predizioni è quindi strettamente connessa alla correttezza dell'assunzione effettuata in partenza.
Nella seconda tipologia, invece, rientrano quei metodi che estraggono conoscenza dalla rete, al fine di poter effettuare previsioni. Rispetto ai modelli non supervisionati, non sono introdotte direttamente dall'utente misure che permettano il calcolo di score di confidenza ma, al contrario, le predizioni emergono da un processo di analisi che tende a estrarre conoscenza dallo stato osservato della topologia. Rispetto ai modelli non supervisionati l'utilizzo di un approccio supervisionato riesce, a dispendio di una maggiore complessità, a garantire migliori performance predittive.

Proprio per garantire migliori performance nell'accuratezza delle predizioni, in questo lavoro è stato scelto di investigare una famiglia di approcci supervisionati: a tale fine abbiamo affiancato un processo ben noto nel Data Mining, la classificazione, ad algoritmi non supervisionati di Link Prediction. Il modo più naturale per utilizzare questa metodologia d'indagine è stato quello di costruire un insieme di classificatori su set di attributi rappresentanti le caratteristiche topologiche della rete. Nel nostro caso sono stati analizzati sei classificatori appartenenti a tre famiglie (Decision Trees, SVM, Bayesiani).

Prima di costruire i classificatori, sono stati scelti e analizzati i dataset utilizzati. La scelta è ricaduta su tre reti sociali non dirette: reti in cui gli archi non hanno un verso e quindi possono essere percorsi in qualsiasi direzione. Nello specifico, le reti analizzate rappresentano sample di noti online social network: Foursquare (un servizio che consente di registrare i

propri spostamenti e condividerli con i propri contatti), Last.fm (una comunità online incentrata sull'ascolto di musica) e Facebook.

I classificatori sono stati costruiti concentrandosi sulla predizione degli archi, adottando due differenti set di feature per ogni rete.
Nel primo caso è stata eseguita una previsione a classi *non filtrate*, in cui le feature sono state calcolate per tutte le potenziali coppie di nodi nella rete. Nel secondo caso, invece, è stata eseguita una previsione a classi *filtrate* in cui è si sono considerate (durante la fase di apprendimento del modello di classificazione) le predizioni solo per le coppie di nodi aventi almeno un vicino in comune. Al fine di rendere l'analisi più agevole ci siamo concentrati nel particolare caso di analisi definito a "*classi bilanciate"*: tutti i modelli di classificazione analizzati sono stati infatti costruiti a partire da un egual numero di istanze positive (feature estratte da coppie di nodi connesse tramite un link) e negative (feature estratte da coppie di nodi non connesse da alcun link).

Una volta decisi i criteri di classificazione da adottare è stata affrontata la parte più critica del lavoro: la scelta delle feature di classificazione. Dato il problema studiato, è stato necessario includere tutte quelle caratteristiche che rappresentassero, direttamente e\o indirettamente, una forma di vicinanza tra i nodi.
Per migliorare i risultati ottenuti, è stata introdotta inoltre un'ulteriore feature atta a catturare il numero di comunità condivise da ciascuna coppia di nodi. Quest'ultima informazione è caratterizzante nelle reti estratte da interazioni sociali poiché, come mostrano numerosi studi, in tali contesti vale una proprietà nota come "*omofilia"*: nodi simili tendono ad essere connessi tra loro con maggiore probabilità e a raccogliersi in gruppi omogenei.
Infine è stata eseguita un'analisi comparativa su i modelli che i classificatori hanno elaborato.

Il problema di Link Prediction è importante e utile perché permette di analizzare e comprendere le dinamiche di gruppi sociali. Un suo studio attento può portare alla creazione di strumenti atti a di identificare legami nascosti e\o identificare connessioni future. Per raggiungere tale fine, in questo lavoro si è mostrato come sia possibile effettuare previsioni aventi un buon grado di confidenza a patto di applicare metodologie di analisi adeguate: gli algoritmi di classificazione adottati si sono, infatti, dimostrati molto efficienti nel predire nuovi archi futuri.

Questo lavoro è stato strutturato in cinque parti.

Nella *Parte I*, dove si analizza lo Stato dell'arte, viene introdotto il modello matematico di grafo, utilizzato durante la trattazione. Sono stati quindi definiti i modelli di rete e spiegate le differenze che ci sono tra di essi. Successivamente, sono state descritte le misure di rete classiche. Una volta definiti i modelli e le misure, sono stati brevemente introdotti alcuni dei problemi comunemente studiati nella teoria dei grafi. Simmetricamente, poiché il lavoro fa uso di modelli di apprendimento supervisionati (i.e. approcci di classificazione), è stato introdotto l'ambito di indagine conosciuto come Data Mining.

Nella *Parte II*, viene introdotto il tema centrale affrontato nel lavoro di tesi. Viene trattato il problema di Link Prediction e, tramite la descrizione delle metodologie di analisi, in particolare degli approcci non supervisionati e supervisionati, vengono definiti i metodi per poter affrontare il problema. Vengono presentate, inoltre, le differenze tra i due modelli.

Nella *Parte III* sono definite le tecniche e gli algoritmi di classificazione usati durante la tesi. Sono qui descritti gli alberi di decisione (dettagliando le caratteristiche degli algoritmi utilizzati appartenenti a questa famiglia, quali C4.5, Bagging e Random Forest), le Support Vector Machine, ed i classificatori Bayesiani (rimandando al teorema di Bayes e alle reti bayesiane, temi fondamentali per questa famiglia, e dettagliando gli

algoritmi usati, quali Naive Bayes e BayesNet).

Nella *Parte IV* viene quindi discussa l'analisi sperimentale effettuata. Sono qui descritte e analizzate le tre reti utilizzate (Facebook, Foursquare, Last.fm); descritta la metodologia di lavoro, i criteri organizzativi seguiti e, in forma tabellare e grafica, riportati e discussi i risultati emersi dai modelli di classificazione adottati.

Infine nelle conclusioni, *Parte V*, si propone una rilettura globale per i risultati ottenuti e viene fornita un'analisi complessiva dell'andamento delle diverse tipologie dei classificatori.

PARTE I
STATO DELL'ARTE

Stato dell'arte

Dato il ruolo importante svolto dai sistemi complessi nella scienza e nell'economia, la loro comprensione, descrizione matematica, previsione e, infine, controllo rappresenta una delle principali sfide scientifiche e intellettuali del ventunesimo secolo. Dietro ogni sistema complesso vi è una rete intricata che codifica le interazioni tra le componenti del sistema.

Negli ultimi anni si è compreso che sistemi anche molto diversi tra loro possono essere efficacemente descritti in termini di "networks" o reti complesse. Esempi ne sono le reti di tipo tecnologico, come Internet o il WWW, le reti di tipo biologico come reti metaboliche o proteiche, e infine reti di tipo sociale, come ad esempio quelle usate per rappresentare le collaborazioni in ambito scientifico o la struttura delle grandi organizzazioni aziendali.

In generale, una rete è una schematizzazione di un sistema costituito da molte entità, dette nodi, (che a seconda del contesto rappresentano persone, computer, proteine ecc.), legate tra loro e interagenti mediante connessioni. In tutti questi casi, la caratteristica comune è l'esistenza di proprietà topologiche complesse.

Lo studio delle caratteristiche strutturali delle reti complesse è di notevole importanza per la comprensione dei meccanismi e delle leggi generali che regolano la diffusione di informazioni, attraverso i nodi della rete, ma anche al fine di comprendere le debolezze di una rete, quelle caratteristiche che la rendono vulnerabile a disturbi esterni.

1. Network science

Ogni rete può essere descritta tramite proprietà strutturali che ne caratterizzano il comportamento. Per comprendere appieno come queste proprietà influenzino un sistema complesso è necessario prendere confidenza con la teoria dei grafi, ramo della matematica che sta alla base della teoria delle reti.

Il modello matematico noto con il nome di "Grafo" è comunemente utilizzato per modellare problemi di vari ambiti di indagine teorica e pratica: la sua formazione è da attribuire al matematico e fisico svizzero Eulero, il quale nel 1736 introdusse questa rappresentazione nella sua pubblicazione sul problema dei "Sette ponti di Konigsberg"(figura 1). Risolvere il problema di Konigsberg significa trovare una strada intorno alla città che permetta di attraversare ciascun ponte soltanto una volta. Eulero sostituì le quattro zone di terra con dei nodi (da A a D) e ogni ponte con un link (da "a" a "g"), ottenendo un grafo con quattro nodi e sette link. Attraverso questa rappresentazione dimostrò che sul grafo di Konigsberg, non può esistere alcun percorso che attraversi ogni link una sola volta.

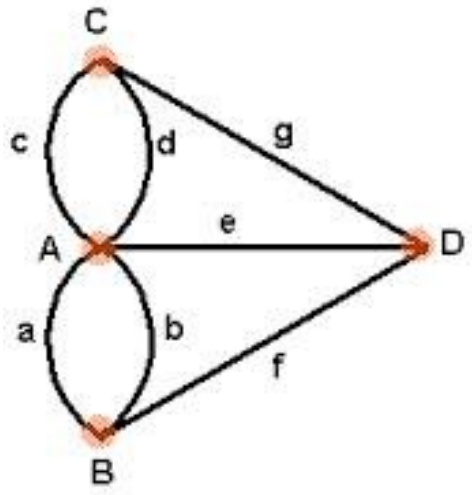

Figura 1: i ponti di Konigsberg

Altri noti problemi affrontati con questo approccio durante il XIX e il XX secolo sono quelli noti come "Il problema dei quattro colori" e "Ciclo Hamiltoniano". Nella seconda metà del XX secolo la teoria dei grafi si è ampliata, grazie allo sviluppo della combinatoria e del calcolo automatico.

Di seguito viene riportata la definizione di Grafo.

Definizione 1. (Grafo)
Si definisce un grafo G come una coppia di elementi (V,E) dove V è l'insieme dei nodi e E è l'insieme degli archi.
Un grafo è *non ordinato,* nel caso in cui E è un insieme di coppie non ordinate di vertici.
Un grafo è *ordinato,* nel caso in cui E è un insieme di coppie ordinate di vertici.

I collegamenti di una rete possono essere orientati o non orientati (figura 2).
Un grafo è detto non orientato se gli archi non hanno un verso e quindi possono essere percorsi in qualsivoglia direzione.
Un grafo è orientato se ciascun arco è caratterizzato da un verso. In particolare, in questo caso, ogni arco è composto di una testa, che raggiunge il vertice in entrata, e da una coda, che lascia il vertice in uscita.

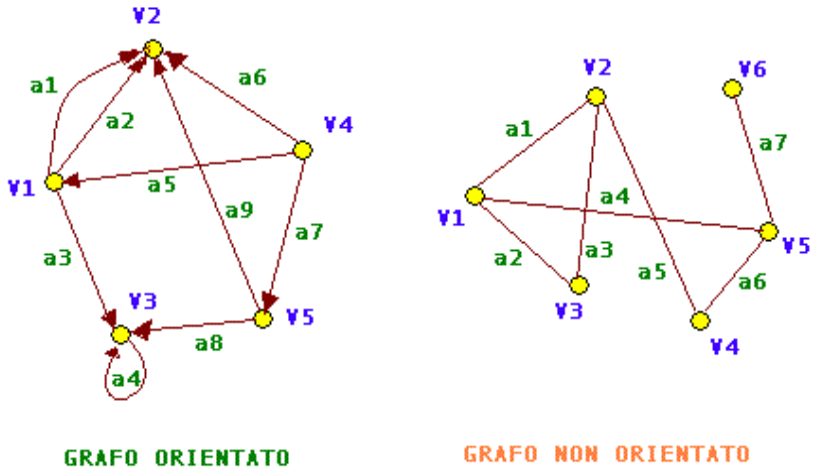

Figura 2: Grafo orientato e grafo non orientato

Questo modello matematico consente l'impiego dei grafi in molti ambiti di ricerca e permette la rappresentazione di diverse tipologie di rete.

Una descrizione completa di una rete ci impone di tenere traccia dei suoi collegamenti. Il modo più semplice per raggiungere questo obiettivo è quello di fornire una lista completa dei link. Per questo viene usata una matrice di adiacenza dove:

- A_{ij} =1, se vi è un link che punta dal nodo j al nodo i;
- A_{ij} =0, se i nodi i e j non sono collegati.

In figura 3 un esempio di matrice di adiacenza.

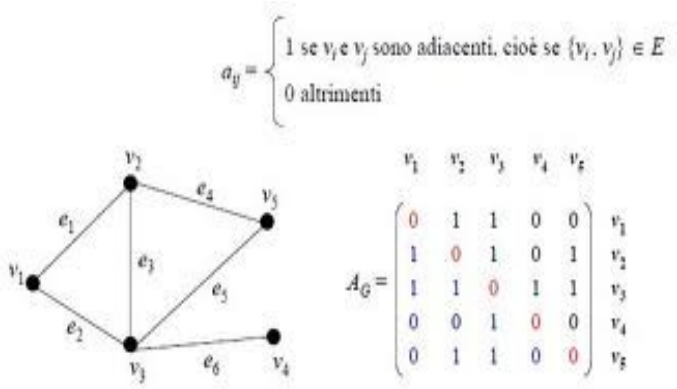

Figura 3: matrice di adiacenza

1.1 Le reti nel mondo reale

Modellando un contesto applicativo su una rete e attribuendo il corretto significato ai componenti di un grafo è possibile suddividere le reti in quattro macro categorie [1].

Reti sociali

Una rete sociale è una struttura composta da individui, organizzazioni o altre entità, inserite in un contesto sociale. Queste entità sono collegate tra loro, da relazioni che possono rappresentare: amicizia, interazione, rapporti di collaborazione o influenza.

Reti tecnologiche

In questa categoria rientrano tutte quelle reti progettate dall'uomo per la distribuzione di beni o servizi. Alcuni esempi possono essere la rete elettrica, la rete di rotte aeree, le reti stradali e ferroviarie, le reti telefoniche e Internet.

Reti di informazione

Rientrano nella categoria delle reti di informazione, chiamate anche reti di conoscenza, le reti costruite sulle citazioni accademiche e la rete del World Wide Web.

A livello topologico le due reti sono diverse.

Le reti di citazioni hanno una struttura che riflette quella delle informazioni memorizzate nei suoi vertici; esse sono acicliche perché i documenti possono citare solo quelli pubblicati.

A differenza di una rete di citazione, il World Wide Web è ciclico; non ha restrizioni sui link data la modificabilità dei testi una volta pubblicati.

Reti biologiche

Sono quelle reti che descrivono sistemi biologici. Alcuni esempi coinvolgono le reti di interazioni tra proteine, quelle costruite su reazioni metaboliche o epidemiologiche e le reti neurali.

1.2 Modelli di reti

Un obiettivo importante della scienza delle reti è quello di costruire modelli che riproducano con precisione le proprietà osservate nei sistemi reali. In 1.2.1 e in 1.2.2 vedremo i modelli più conosciuti e in 1.2.3 saranno spiegate le differenze tra i modelli.

1.2.1 Random Graph

I grafi random (figura 4) rappresentano il modello più semplice per le reti complesse. In questa maniera si arriva alla definizione di rete random [4]:

Dato un numero fisso di n vertici, gli archi vengono creati in modo casuale, secondo una certa distribuzione di probabilità.

Esistono due modi equivalenti per definire le reti casuali:

il primo è dovuto allo studio fatto da Erdos e Renyi nel 1959. La costruzione di un grafo random G(N,L), dove i nodi N etichettati sono collegati con L collegamenti posizionati casualmente, parte dalla condizione iniziale di N nodi privi di connessioni. Il grafo è generato inserendo un arco tra coppie di nodi scelti in maniera casuale, impedendo

connessioni multiple tra la stessa coppia di nodi e ripetendo l'operazione fino a quando il numero di archi raggiunge la quantità prestabilita L.
Il secondo è il modello introdotto da Gilbert nel 1959; esso è composto da un grafo G(N,p), in cui ogni possibile arco tra due vertici è presente con identica probabilità p, e assente con probabilità 1-p.
I grafi random sono caratterizzati da buone qualità globali; infatti, la distanza media tra due nodi aumenta lentamente all'aumentare di N, e risulta abbastanza piccola anche in reti con un elevato numero di nodi. Di contro, questo modello differisce dalle reti reali per due motivi principali:
il primo motivo è che la distribuzione del grado segue un'irreale legge di Poisson; il secondo è che in tale modello, le reti hanno un basso coefficiente di clustering.

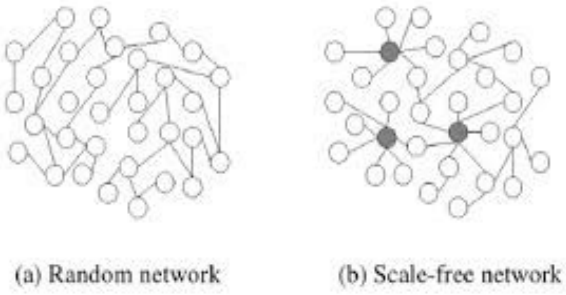

Figura 4: modelli di rete

1.2.2 Scale – free networks

Studiando la struttura del web, attraverso un web crawler (un programma capace di navigare il web in maniera metodica per raccogliere informazioni), il fisico Albert Barabasi [2,4] scoprì che questa non presentava una connettività casuale. Notò, invece, la presenza di alcuni nodi con un numero di archi elevato, che vennero denominati hub o connettori. Gli hub hanno una grandissima importanza all'interno di una rete; dominano la struttura di tutte le reti in cui sono presenti rendendole simili a mondi piccoli. Infatti, gli hub, essendo collegati ad un numero insolitamente grande di nodi, accorciano tutte le distanze all'interno del sistema. Quindi, gli hub rappresentano una caratteristica importante di sistemi complessi anche molto diversi tra loro. La scoperta dell'esistenza dei connettori ha portato la necessità di un nuovo modello che tenesse conto dell'esistenza di questo particolare tipo di nodi.

Le reti ad invarianza di scala (figura 4), o scale-free networks, presentano una distribuzione di grado di tipo power law (legge di potenza) e che meglio si prestano per studiare la presenza degli hub e i loro effetti. Il modello maggiormente utilizzato per la generazione di modelli atti a studiare reti ad invarianza di scala è quello proposto da Barabasi e Albert [2,4] nel 1999, denominato “i ricchi sono sempre più ricchi”. La nascita di una rete a invarianza di scala è molto semplice: quando un nodo deve stabilire un nuovo collegamento, si assume che preferisca farlo verso un nodo che ne ha già molti, portando questi ultimi ad una crescita esponenziale delle proprie connessioni con l'aumentare del numero dei collegamenti della rete (figura 5).

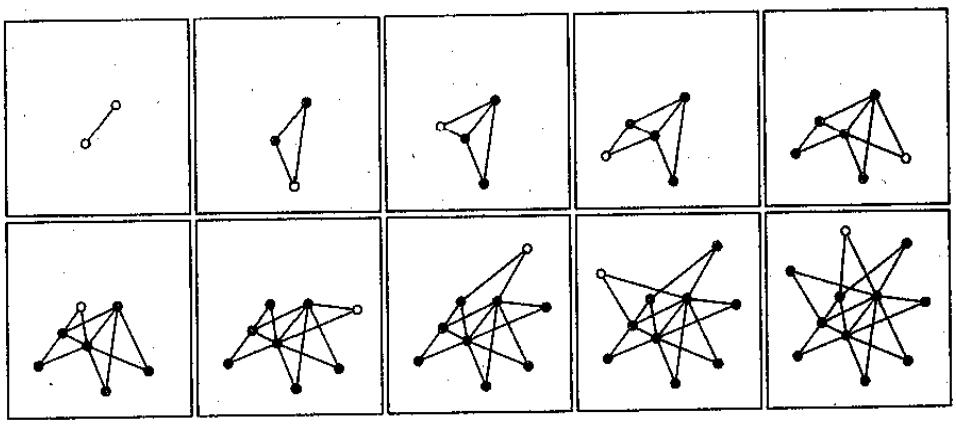

Figura 5: Nascita di una rete a invarianza di scala

Definizione 2. (Power Law)

2 variabili x e y sono legate da una legge di potenza quando:

$$y(x) = Ax^{-\gamma}$$

dove A e γ sono costanti positive. La costante γ è spesso chiamata esponente della legge di potenza.

Molti fenomeni naturali e sociali sono caratterizzati dalla presenza di una scala intrinseca.

Si parla di scala intrinseca quando le distribuzioni dei valori numerici delle variabili quantitative tipiche dei fenomeni studiati tendono a concentrarsi in un intervallo ben definito, mentre i valori lontani da tale intervallo sono estremamente improbabili.

Esistono tuttavia, sia in natura sia nell'ambito delle scienze umane, numerosi fenomeni che non possiedono una scala intrinseca.

La caratteristica più evidente delle distribuzioni che descrivono il comportamento quantitativo dei sistemi privi di scala sta nel fatto che le distribuzioni non si concentrano in intervalli particolari, ma nelle situazioni

realistiche ci saranno valori “grandi” che compaiono con scarsa frequenza e valori “piccoli” che compaiono con maggior frequenza.

In mancanza di una scala intrinseca la riduzione della frequenza al crescere del valore tenderà ad avvenire con la stessa regolarità in tutti gli intervalli di valori, per cui a una determinata variazione percentuale (positiva) del valore corrisponderà in ogni intervallo una costante e determinata variazione percentuale (negativa) della frequenza.

Questo tipo di relazione si traduce in una proprietà matematica ben precisa delle distribuzioni, che prendono la forma di leggi di scala o di potenza [12].

In figura 6 è riportata la distribuzione di una legge di potenza.

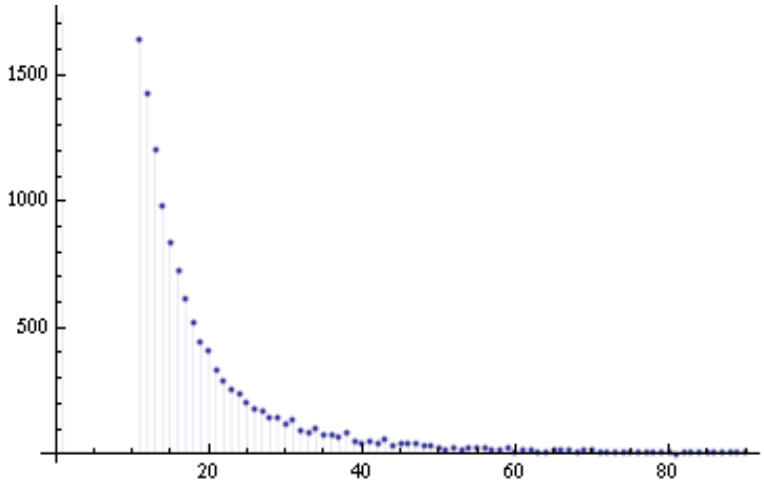

Figura 6: Distribuzione di una legge di potenza

Il primo che individuò una distribuzione che obbediva ad una legge di potenza fu Pareto. La legge 80/20 da lui scoperta è sintetizzabile nell'affermazione: la maggior parte degli effetti è dovuta ad un numero ristretto di cause.

1.2.3 Differenza tra i due modelli

Il modello random di Erdos e Renyi è basato su due semplici assunti. In primo luogo, il numero dei nodi è fissato fin dall'inizio, e rimane invariato per l'intera durata di vita della rete. In secondo luogo, i nodi si equivalgono. Di conseguenza, le reti generate da questo modello sono statiche.

Le reti proposte da Barabasi, invece, sono governate da due leggi:

Crescita continua: in ogni intervallo di tempo un nuovo nodo si aggiunge.

Collegamento preferenziale: assumendo che ogni nuovo nodo si connetta necessariamente ad altri già presenti una volta comparso, la probabilità che scelga un certo nodo è proporzionale al numero di link da questi posseduto.

Questo modello, combinando la crescita e il collegamento preferenziale, è stato il primo tentativo di spiegare gli hub e le leggi di potenza e venne definito modello ad invarianza di scala.

Attraverso la crescita, la rete si espande e questo comporta che i primi nodi abbiano, rispetto agli ultimi, più tempo per acquisire un link. Al nodo arrivato per ultimo non potrà collegarsi nessun altro nodo; al primo nodo della rete, invece, potranno collegarsi tutti i nodi successivi. Ciò comporta che la crescita offre un notevole vantaggio ai nodi più vecchi che diventano sempre più ricchi.

Attraverso il collegamento preferenziale, i nuovi nodi preferiscono connettersi con quelli più ricchi di link; i nodi più vecchi verranno selezionati più frequentemente e cresceranno più rapidamente rispetto agli altri. I nuovi nodi che arriveranno continueranno a scegliere i nodi più connessi , i quali accumuleranno un numero altissimo di link e si distanzieranno dalla media, trasformandosi in hub. Il collegamento preferenziale porta ad un fenomeno in cui i “i ricchi diventano sempre più ricchi”. Ciò conduce alle leggi di potenza osservate nelle reti del mondo reale.

Le due leggi alla base dell'evoluzione della rete, strutturate nel modello a invarianza di scala, offrono un buon punto di partenza per esplorare questi diversi sistemi.

2. Misure di rete

Definiti i principali modelli delle reti, in questo capitolo vengono introdotte alcune importanti caratteristiche.

2.1 Fenomeno dello Small world

Una delle caratteristiche osservate nelle reti sociali è lo Small World, il mondo piccolo, che può essere interpretato come la dimostrazione che le persone sono collegate tra loro da una piccola catena di conoscenti.

Nel 1929, per la prima volta nella storia, lo scrittore ungherese Frigyes Karinthy, in un breve racconto intitolato *"Chains"*, espose la teoria dello Small World. Nel 1967 il sociologo Stanley Milgram [10] decise di verificare la teoria di Karinthy tramite un esperimento. Milgram ipotizzò dunque che i membri di una qualsiasi rete sociale fossero interconnessi tra loro tramite una breve catena di conoscenze intermedie. Milgram procedette nel seguente modo: selezionò casualmente 296 persone (residenti in Nebraska) a cui inviò una lettera contenente il nome della persona (ovvero un agente di Boston) a cui la lettera doveva essere recapitata e la richiesta di aggiungere alcune informazioni personali e rispedire la lettera ad un conoscente che, basandosi solo sul nome del destinatario, sul fatto che abitasse a Boston e sul suo lavoro, potesse essere in grado di contattarlo o comunque avesse maggiori possibilità di far arrivare la lettera a destinazione.

Figura 7: esperimento di Milgram

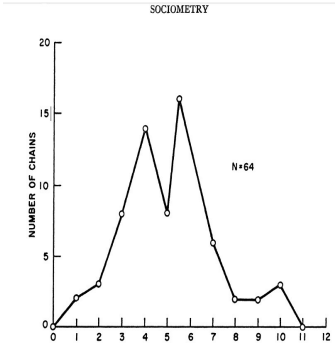

Figura 8: Grafico della prova

Nella figura 8 si visualizza la distribuzione delle lunghezze dei cammini completati, ovvero le lettere che dal Nebraska sono giunte a destinazione. La lunghezza di una catena completata è definita come il numero di intermediari richiesti per collegare il mittente con il destinatario. La media della distribuzione è di 5.2 passaggi (links).

Nel 2001, Duncan Watts [11], un professore della Columbia University, ricreò l'esperimento di Milgram sfruttando Internet come mezzo. Watts usò un messaggio di posta elettronica come "pacchetto" che doveva essere consegnato e, sorprendentemente, dopo aver analizzato i dati ottenuti dagli invii effettuati da 48.000 differenti persone residenti in 157 stati diversi, nei confronti di 19 "bersagli", Watts riscontrò che il numero medio di intermediari era nuovamente 6, riconfermando la validità dell'esperimento di Milgram. A differenza di Milgram il target di riferimento conta 18 persone (12 maschi e 6 femmine) aventi tipi di occupazioni diverse e di 13 nazionalità diverse. E' interessante notare che, per far giungere la mail a destinazione, sia il 57% dei maschi che il 61% delle femmine hanno inviato il messaggio ad altri utenti dello stesso sesso.

Da notare, inoltre, che i mittenti hanno scelto di mettersi in contatto con il destinatario preferenzialmente tramite una rete di amicizie rispetto alla rete familiare o di business.

La ricerca di Watts e l'avvento dell'era di Internet ha permesso di applicare la teoria dei sei gradi di separazione anche in ambiti differenti, tra cui l'analisi delle reti informatiche, delle reti elettriche, la trasmissione di malattie, le telecomunicazioni e la progettazione della componentistica per apparati elettronici.

2.2 Coefficiente di clustering

La proprietà di clustering, o transitività, di un grafo misura il grado di "compattezza" di una rete, ossia la tendenza di due nodi, adiacenti ad un nodo comune, di essere connessi l'uno all'altro.

Esso viene definito come:

$$C_i = \frac{2L_i}{k_i(k_i - 1)}$$

dove L_i rappresenta il numero di collegamenti tra i vicini K_i del nodo i. Il coefficiente di clustering è compreso tra 0 e 1:

- C_i= 0, se nessuno dei vicini del nodo i è collegato ad altri;
- C_i=1, se tutti sono collegati.

In generale, il coefficiente di clustering misura la densità locale della rete; più sono densamente interconnessi i vicini del nodo i, maggiore è il coefficiente di clustering.

Oltre a questa versione locale, esiste una formulazione globale di coefficiente di clustering [1,4]. Essa tende a dare informazioni relative al grado di clustering della rete.
Il coefficiente di clustering globale analizza triple di nodi e misura il numero totale di triangoli chiusi in una rete.
E' definito come:

$$C = \frac{3 * \text{numero di triangoli}}{\text{numero di triple di vertici connesse}}$$

dove con triple connesse si intende un singolo vertice con archi che conducono ad un altro paio di vertici.

2.3 Componente gigante e componenti connesse

In una rete non orientata due nodi i e j appartengono alla stessa componente se esiste un cammino tra di loro sulla rete. Essi sono scollegati se tale percorso non esiste; in questo caso si ha $d_{ij} = \infty$ (dove d sta per distanza). Ciò è illustrato nell'immagine 9a, che mostra una rete composta da due cluster disconessi. Mentre ci sono percorsi tra i nodi che appartengono allo stesso cluster (ad esempio nodi 4 e 6), non ci sono percorsi tra i nodi appartenenti a diversi cluster (nodi 1 e 6).
Una rete è connessa se esistono cammini che connettono tutte le coppie di nodi nella rete. Essa è disconnessa se esiste almeno una coppia di nodi con $d_{ij} = \infty$.

Una componente è un sottoinsieme di nodi in una rete per cui esiste un percorso tra ogni coppia di nodi che ne fanno parte. Se la rete è costituita da due componenti, un singolo collegamento posizionato correttamente può collegarli, rendendo la rete connessa (figura 9b). Tale collegamento è chiamato ponte. Un ponte è un legame che, se tagliato, disconnette il grafo.

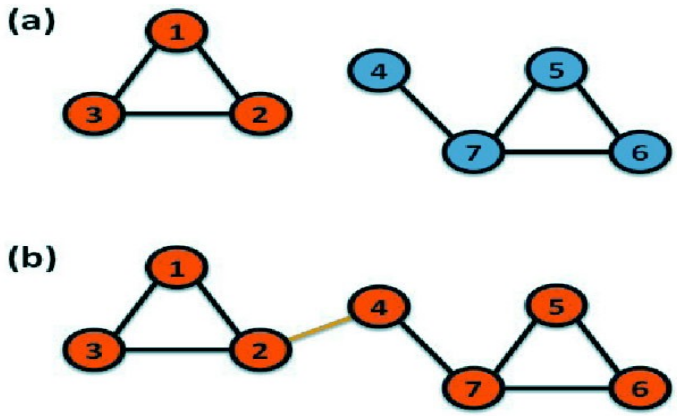

Figura 9: Reti connesse e reti sconnesse

Durante l'esplorazione di un grafo può, spesso, essere utile tenere traccia della lunghezza del cammino percorso o comunque avere un'unità di misura per quantificare il numero di archi osservati. Per soddisfare questa nuova necessità vengono introdotte le definizioni di: shortest path, eccentricità, raggio e diametro.

Shortest path

Un cammino è un percorso che corre lungo le maglie della rete; la sua lunghezza è definita dal numero di collegamenti che esso attraversa. Un percorso può intersecare se stesso e passare attraverso lo stesso collegamento più volte.

Shortest path tra due nodi è il percorso tra due nodi u e v di lunghezza minore. È anche detto distanza geodetica (figura 10).

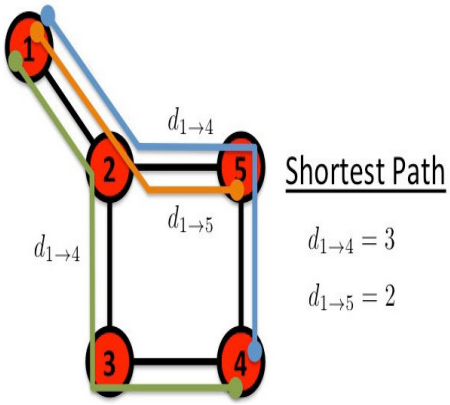

Figura 10: Shortest path

Un'importante proprietà della rete è la lunghezza media dei percorsi più brevi, nota anche come distanza media, definita come la media dei percorsi più brevi , calcolata su tutte le coppie di nodi della rete:

$$\langle d \rangle = \frac{1}{N(N-1)} \sum_{i,j=1,N} d_{i,j}$$

E' auspicabile che una rete abbia un valore di <d> che sia il più basso possibile, poiché ciò implica che la rete soddisfa la proprietà di small world.

Eccentricità

Dato un grafo G=(V,E), viene definita eccentricità di un vertice v, la più grande distanza tra v e ogni altro vertice di G.

Raggio

Dato un grafo G=(V,E), si definisce raggio del grafo la minima eccentricità tra tutti i vertici.

Diametro

Dato un grafo G=(V,E), si definisce diametro, indicato con dmax, il massimo cammino minimo nella rete.

2.4 Misure di centralità

La centralità è un indicatore in grado di motivare l'importanza di un nodo in una rete.

Nello studio delle reti complesse la centralità può essere importante per:

- giudicare la rilevanza o la criticità di nodi o aree della rete;
- attribuire una misura di distanza fra nodi o aree della rete;
- identificare il grado di coesione di un'area della rete;
- identificare le aree di una rete.

Le misure di centralità più importanti sono: il grado, la closeness, la betweenness e l'eigenvector centrality.

Grado(Degree)

Una proprietà chiave di ciascun nodo è il suo grado, che rappresenta il numero di archi ad esso incidenti [4].

In una rete non diretta, identificata da un grafo non orientato G=(V,E), si definisce grado di un vertice v il numero di archi ad esso incidenti. Intuitivamente questa grandezza fornisce una misura dell'importanza del nodo all'interno del grafo.

Viene indicato con K_i il grado del nodo i-esimo della rete.

In una rete non diretta il numero totale di link L, può essere espresso come la somma del grado dei nodi:

$$L = \frac{1}{2}\sum_{i=1}^{N} k_i$$

Un importante proprietà della rete è la media del grado nelle reti indirette:

$$\langle k \rangle \equiv \frac{1}{N}\sum_{i=1}^{N} k_i = \frac{2L}{N}$$

In una rete diretta, dato un grafo orientato G=(V,E), si definisce grado entrante (uscente), o in-degree (out-degree), di un vertice v che appartiene a V, il numero di archi entranti (uscenti) da esso. Il grado di un vertice v, in

un grafo orientato, è dato dalla somma del suo grado entrante e del suo grado uscente:

$$k_i = k_i^{in} + k_i^{out}$$

Il totale numero di links in una rete diretta è:

$$L = \sum_{i=1}^{N} k_i^{in} = \sum_{i=1}^{N} k_i^{out}$$

Il grado medio di una rete diretta è:

$$\langle k^{in} \rangle = \frac{1}{N} \sum_{i=1}^{N} k_i^{in} = \langle k^{out} \rangle = \frac{1}{N} \sum_{i=1}^{N} k_i^{out} = \frac{L}{N}$$

La distribuzione di grado, p_k , fornisce la probabilità che un nodo selezionato casualmente nella rete abbia grado K.

Closeness Centality

La closeness centrality fornisce una misura della distanza di un nodo da tutti gli altri nodi. Al contrario della degree centrality, questa metrica necessita di una visione globale della rete e non è quindi limitata alla visione locale dei singoli nodi.

Per ottenere valori elevati per le piccole somme di distanze, viene calcolata come l'inverso della distanza totale:

$$C_C(v) = \frac{1}{\sum_{t \in V} d_G(v,t)}$$

Così la distanza da un vertice con alta closeness ad un altro è breve in media. Questi vertici sono considerati strutturalmente importanti perché possono raggiungere o essere raggiunti da altri nodi.

Betweenness Centrality

Un concetto alternativo di centralità è basato sull'idea del controllo sulle connessioni tra coppie di vertici.

La Betweenness centrality si basa sull'osservazione che se un nodo fa parte di molti cammini minimi allora e un nodo che riveste una posizione importante nel dominio del sistema che stiamo rappresentando come rete.

La formula è la seguente:

$$c_B(v) = \sum_{s,t \in V} \frac{\sigma(s,t|v)}{\sigma(s,t)}$$

Eigenvector Centrality

L'eigenvector centrality misura la centralità di un nodo in base alle sue interazioni con la rete. Un nodo è importante se collegato ad altri nodi importanti.

Il pagerank [5] è basato su questa misura di centralità.

2.5 Densità

La densità misura la percentuale di collegamenti presenti in un grafo rispetto al totale dei possibili. Vi sarà maggiore densità quanto più numerosi saranno i collegamenti tra i nodi.

La densità dipende da:

- l'inclusività, cioè dal numero totale dei nodi collegati meno i nodi isolati;
- dal grado di connessione: alcuni nodi avranno collegamenti con molti altri, mentre altri nodi avranno collegamenti meno numerosi. Quanto più nodi hanno elevate connessioni, tanto più il grafo sarà denso.

Nei grafi non orientati la densità viene così calcolata:

$$densità = \frac{l}{n(n-1)/2}$$

dove l= numero di archi e n= numero dei nodi.

Nei grafi diretti la densità viene così calcolata:

$$l/n(n-1)$$

La densità di un grafo varia tra 0 e 1.

2.6 Resilience

La resilience di un grafo è la misura della sua robustezza, in caso di guasti o rimozioni di nodi o di archi. Se i vertici di una rete vengono rimossi, la lunghezza dei percorsi aumenta e coppie di vertici saranno potenzialmente relegate a componenti non connesse e come conseguenza non sarà possibile la comunicazione tra di loro nella rete. I vertici possono essere rimossi casualmente o colpendo i nodi con il più alto grado.

Le reti variano il livello di resilience alla rimozione dei nodi.

Molti grafi ottenuti da reti reali dimostrano un'alta resilience relativamente a fallimenti randomici di nodi e archi ma bassa resilience in caso di attacchi mirati.

Questi risultati si spiegano grazie alla Power Law che accomuna molte reti del mondo reale.

Attacchi mirati ai pochi nodi aventi altissimo degree in una rete scale-free possono causare una frammentazione della rete in più componenti sconnesse tra loro. Fallimenti casuali procurano pochi danni alla struttura globale della rete.

3. Problemi Complex Networks

Definiti i modelli e le misure di rete e di centralità, passiamo ad introdurre tre problemi comunemente studiati in teoria dei grafi:

- Node Ranking (3.1);
- Community discovery (3.2);
- Link prediction (3.3).

3.1 Node Ranking

Questo problema ha come fine ultimo quello di sfruttare la struttura di un grafo per ordinarne i nodi secondo un criterio di importanza.

Alcuni esempi ben noti si trovano nell'ambito dell'information retrival con gli algortimi PageRank e Hits.

Consideriamo come esempio la rete del web. Questa definisce un grafo orientato G, con le pagine web che definiscono i nodi da 1 a N e i collegamenti tra di esse che definiscono gli archi.

Questo grafo può essere descritto da una matrice di adiacenza A, dove

A_{ij} = 1, se vi è un link dalle pagina i alla pagina j, e A_{ij} =0, se non vi è nessun link dalla pagina i alla pagina j.

Una nozione importante trattata nell'articolo [5] che è al centro delle descrizioni degli algortimi PageRank e Hits è quella di eigenvector, vista in 2.4.

PageRank

PageRank, sviluppato da Brian e Page [5,13,14], è un algoritmo di link analysis che assegna un peso numerico ad ogni documento ipertestuale, con lo scopo di misurare la sua importanza all'interno della rete. Il peso che viene assegnato ad un documento u è indicato come PageRank di u.

Hits

L'algoritmo Hits è stato sviluppato da Kleimberg [5] e viene usato per la ricerca di collezioni di pagine autorevoli per le cosiddette broad topic query, interrogazioni su argomenti di carattere generale.

Il modello di Hits considera le relazioni esistenti tra le pagine autorevoli, denominate autorithy, per un'interrogazione di carattere generale, e quelle che presentano link verso numerose autorithy, dette hub; l'algoritmo identifica e sfrutta il valore informativo di entrambe le tipologia di pagina.
L'obiettivo è quello di estrarre dal bacino di pagine su cui l'algoritmo viene applicato, quelle ad elevato punteggio di autorithy sfruttando, a tal fine, le indicazioni di autorevolezza fornite dalle pagine con elevato punteggio di hub.
In figura 11 uno schema dell'algoritmo.

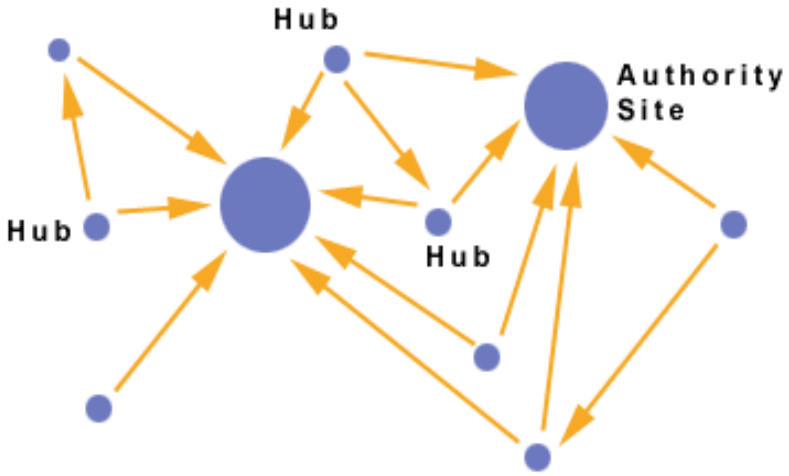

Figura 11: hub autorithies

3.2 Community Discovery

Un problema critico che è stato ampiamente studiato negli ultimi anni, è l'identificazione delle comunità nascoste all'interno della struttura di reti complesse. L'obiettivo del community discovery è quello di clusterizzare i nodi appartenenti ad un grafo che condividano determinate caratteristiche [6].
Una comunità è intesa come un insieme di entità, dove ogni entità è più vicina, nel senso di rete, ad altri soggetti all'interno della comunità rispetto alle entità fuori di essa. Pertanto, le comunità sono gruppi di entità che presumibilmente condividono alcune proprietà e svolgono ruoli simili all'interno del fenomeno interagente che viene rappresentato. Il rilevamento di comunità è importante per molte ragioni, tra cui la node classification.

Nel grafo del web, ad esempio, le comunità possono corrispondere a gruppi di pagine che trattano simili argomenti o, in reti sociali, gruppi di relativi individui strettamente connessi.
In reti dinamiche, reti che evolvono nel tempo, le comunità possono mutare la loro struttura e composizione.

Demon

Demon è un algoritmo per la scoperta di comunità.
Esso utilizza un approccio basato sull'estrazione di ego networks, l'insieme dei nodi connessi ad un certo nodo u. Le comunità vengono estratte utilizzando un approccio bottom- up. Ogni nodo dà la prospettiva delle comunità che lo circondano e tutti i diversi punti di vista, poi, vengono fusi insieme in una struttura sovrapposta.
Il risultato di questa combinazione è un insieme di moduli sovrapposti, l'ipotesi delle comunità reali
nel sistema globale, non fatta da un osservatore esterno, ma dagli attori della rete stessa.

3.3 Link Prediction

Le reti sociali sono oggetti altamente dinamici; crescono e cambiano rapidamente nel tempo attraverso l'aggiunta di nuovi archi. Il problema di prevedere nuove connessioni nella rete è definito link prediction.
Partendo dalla conoscenza fornita da uno snapshot della rete, ad un prestabilito istante t, è interessante poter predire quali saranno gli archi che si aggiungeranno alla rete in un intervallo di tempo t' (figura 12)[7]. L'obiettivo che ci si pone è quello di predire come si evolverà la struttura della rete analizzata sfruttando le informazione topologiche da essa stessa fornite.

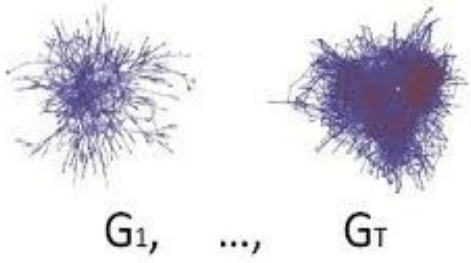

Figura 12: Link Prediction

4. Data mining

Per data mining si intende l'applicazione di una o più tecniche che consentono l'esplorazione di grandi quantità di dati, con l'obiettivo di individuare le informazioni più significative e di renderle disponibili e direttamente utilizzabili nell'ambito scientifico, sociale e del marketing.

L'estrazione di conoscenza, ossia di informazioni significative, avviene tramite individuazione delle associazioni, patterns, o sequenze ripetute, nascoste nei dati. In questo contesto un pattern indica una struttura, un modello, o, in generale, una rappresentazione sintetica dei dati.

Il termine data mining è utilizzato come sinonimo di knowledge discovery in databases (KDD), anche se sarebbe più preciso parlare di knowledge discovery quando ci si riferisce al processo di estrazione della conoscenza, e di data mining come di una particolare fase del suddetto processo.

Questa disciplina trae ispirazioni dalle aree del machine learning, dell'intelligenza artificiale, del pattern recognition, della statistica e di basi di dati.

Un processo di estrazione di conoscenza percorre alcune fasi che possono essere schematizzate in:

- definizione dell'obiettivo;
- individuazione delle fonti di dati;
- estrazione ed acquisizione dei dati;
- pre-processing;
- data mining;
- interpretazione e valutazione dei risultati;
- rappresentazione dei risultati.

La figura 13 mette in luce la natura iterativa del processo. La fase di valutazione può, infatti, portare da una semplice ridefinizione dei parametri di analisi utilizzati, ad una ridefinizione dell'intero processo a partire dai dati estratti.

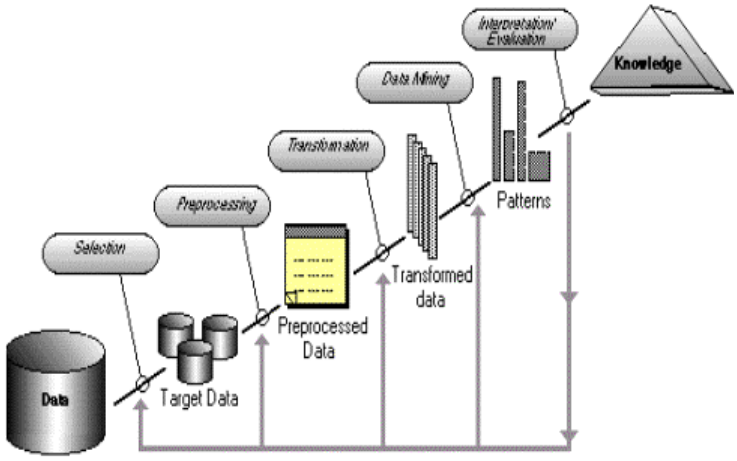

Figura 13: processo di estrazione di conoscenza

Tra le tecniche utilizzate nel data mining vi sono:

- classificazione;
- cluster;
- regole associative;
- regressione.

4.1. Classificazione

Data una collezione di record (training set), dove ogni record è composto da un insieme di attributi di cui uno esprime la classe di appartenenza del record, la classificazione è un procedimento atto a determinare un profilo descrittivo per ogni classe, che permetta di assegnare oggetti di classe ignota alla classe appropriata.

La costruzione del classificatore si basa su un insieme di osservazioni di stima o di training di cui è già noto il valore della classe: a partire da tali esempi, il modello apprende le regole che determinano l'appartenenza ad una certa classe.

L'obiettivo è quello che record non noti siano assegnati a una classe nel modo più accurato possibile. Normalmente, il data set fornito è suddiviso in training e test set. Il primo è utilizzato per costruire il modello, il secondo per validarlo.

I classificatori possono essere utilizzati sia a scopo predittivo sia a scopo descrittivo.
Alcune tra le principali tecniche di classificazione sono gli alberi o decision tee, le regole di decisione, le reti bayesiane, le reti neurali, il support vector machines e nearest- neighbor [9].

4.2 Clustering

Il clustering è un insieme di tecniche di analisi dei dati volte alla selezione e al raggruppamento di elementi omogenei in un insieme di dati. Le tecniche di clustering si basano su misure relative alla somiglianza tra gli elementi. In molti approcci questa similarità, o meglio, dissimilarità, è concepita in termini di distanza in uno spazio multidimensionale. La bontà delle analisi ottenute dagli algoritmi di clustering dipende molto dalla scelta della metrica e quindi da come è calcolata la distanza. Gli algoritmi di clustering raggruppano gli elementi sulla base della loro distanza reciproca, e quindi l'appartenenza o meno ad un insieme dipende da quanto l'elemento preso in esame è distante dall'insieme stesso.
Le tecniche di clustering si suddividono in partizionali e gerarchiche.

4.3 Regole associative

Questa tecnica è basata sulla ricerca di relazioni e dipendenze tra determinati attributi delle osservazioni. L'obiettivo non consiste nel prevedere un certo valore, ma informare l'utente della presenza di particolari affinità tra determinati attribuiti, valide per un sottoinsieme significativo dei dati. L'applicazione più importante di tale tecnica è la market basket analysis, il cui scopo è la ricerca di associazioni tra i prodotti acquisiti dai clienti in un supermercato: la scoperta di tali relazioni diventa strategica all'interno del supermercato stesso e si concretizza ad esempio nella pianificazione di campagne promozionali o nella particolare disposizione dei reparti merceoligici.

4.4 Regressione

La regressione è specializzata nella previsione di variabili quantitative. Molti risultati provengono dalla statistica, ambito in cui il problema della regressione è stato ampiamente studiato: il concetto su cui basa tale tecnica si poggia sulla possibilità di approssimare la vera funzione che ha generato le osservazioni. La previsione è verificata sugli scarti di errore generati dal modello rispetto ai veri valori.

PARTE II

LINK PREDICTION

In questa sezione viene introdotto l'argomento centrale affrontato nel lavoro di tesi.
Dopo aver introdotto la definizione di link predicition nel capitolo stato dell'arte (3.3), sarà formulato il problema di Link Prediction su reti dirette ed indirette. Infine verranno trattati i due modelli di approccio al problema: l'analisi unsupervised e quella supervised.

1. Link Prediction

Nell'ambito delle numerose ricerche sulle grandi reti complesse e le loro proprietà, molta attenzione è stata dedicata all'analisi computazionale della struttura delle reti sociali i cui nodi rappresentano persone o altri soggetti incorporati in un contesto sociale e gli archi rappresentano l'interazione, la collaborazione o l'influenza tra entità.
La maggiore disponibilità di grandi insiemi di dati dettagliati che codificano tali reti ha stimolato un ampio studio sulle proprietà fondamentali e l'identificazione delle ricorrenti caratteristiche strutturali.

Le reti sociali sono oggetti altamente dinamici; crescono e cambiano rapidamente nel tempo attraverso l'aggiunta di nuovi archi, che stanno a significare la comparsa di nuove interazioni nella struttura sociale.

Il problema di prevedere nuove connessioni nella rete è definito Link Prediction [7], a cui si è fatto riferimento già nel capitolo stato dell'arte (3.3). L'obiettivo di tale task quello di predire come evolverà la struttura della rete analizzata sfruttando le informazioni topologiche da essa stessa fornite.

Data la sua formulazione, il problema del Link Prediction rientra nell'ambito di indagine che va sotto il nome di Link Mining. Per Link Mining [15] si intende il raggruppamento di tutte le tecniche di data mining che considerano esplicitamente le informazioni fornite dai link appartenenti ad una rete per costruire modelli descrittivi, predittivi e generativi che possano analizzare i dati rappresentati. Il Link Mining nasce dall'intersezione di vari valori su :

- Link Analysis: tecnica di analisi di dati utilizzata per valutare le relazioni tra i nodi;
- Web Mining: applicazione di tecniche di data mining per scoprire i modelli dal web;
- relational learning [16]: tecnica che si basa sulle idee della teoria della probabilità e della statistica per affrontare l'incertezza, incorporando strumenti di logica, database e linguaggi di programmazione per rappresentare la struttura;
- programmazione induttiva: sottoarea dell'apprendimento automatico che rappresenta la sua confluenza con la programmazione logica;
- graph mining: tecnica che riguarda lo studio della struttura dei grafi.

Tutti gli approcci risolutivi per il problema di Link Prediction sono legati alla ricerca del modello evolutivo della rete: quanto più il predittore proposto riuscirà a sfruttare le informazioni topologiche della rete, tanto più la sua precisione sarà accurata e il modello evolutivo da esso descritto sarà simile a quello che regola la rete.

Numerosi approcci al problema di Link Prediction hanno proposto misure per valutare la "vicinanza" tra i nodi appartenenti ad una rete: molte di queste misure sono originate da tecniche mutuate dalla teoria dei grafi e dalla Social Network Analysis.

1.1 Definizione del Problema

Sia $G_0 = (V,E_0)$ un grafo non orientato definito nelle sue componenti di nodi ed archi osservato ad un dato istante t_0. Il task di Link Prediction, dato un istante $t_1 > t_0$ e il relativo insieme degli archi E_1, consiste nel prevedere gli archi che entreranno a far parte del grafo originario nell'istante futuro t1 (archi appartenenti all'insieme $E_{new} = E_1 - E_0$). Per ogni arco nell'insieme dei risultati deve inoltre essere presente uno score di confidenza.

Secondo questa formulazione è necessario che, per ogni arco predetto, sia calcolato un valore che indichi l'attendibilità della predizione. Il risultato dell'applicazione di un predittore è quindi un insieme di triple (nodo, nodo, score), dove lo score sarà diverso da zero per tutti quegli archi che sono attesi a comparire nell'istante futuro.

A tale fine, sono stati proposti numerosi criteri per stimare le probabilità degli archi sia su reti dirette sia su reti indirette: basandosi su una matrice di adiacenza W possiamo quindi restringere i vincoli delle misure di similarità in base alla tipologia di reti analizzate.

Link Prediction su reti dirette

Il presupposto chiave per prevedere nuovi collegamenti su reti dirette è che se due coppie di nodi sono simili tra loro, la probabilità di collegamenti in queste due coppie sarà simile.

In figura 14, P_{ij} e $P_{i'j'}$ sono simili se il nodo i è simile a i' e il nodo j è simile a j'. Per le reti dirette, quindi, si misura la similarità tra le coppie di nodi (i,i') e (j,j').

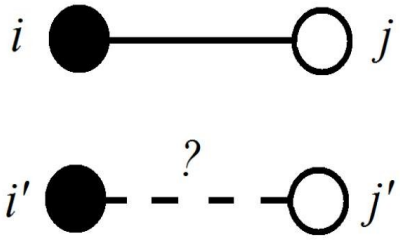

Figura 14: Coppia di somiglianza per la rete diretta

Link Prediction su reti non dirette

Per i grafi non diretti, invece, il presupposto chiave è che P_{ij} e $P_{i'j'}$ sono vicini se due coppie (i,i') e (j,j') sono simili, tenendo conto che la direzione dei due archi non ha importanza. Quindi, le coppie sono simili se sia i è simile a i' e sia j è simile a j', o, se sia i è simile a j' e j sia simile a i' (figura 15)

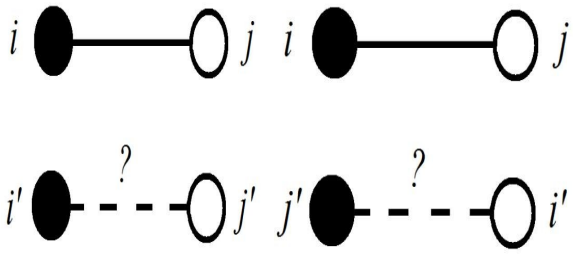

Figura 15: Coppia di somiglianza per la rete indiretta

2. Approcci al problema di Link Prediction

In letteratura il problema di Link Prediction è stato affrontato tramite differenti metodologie di analisi [7,17]. In particolare tutti gli approcci proposti possono essere ricondotti a due categorie:

metodi non supervisionati;

metodi supervisionati.

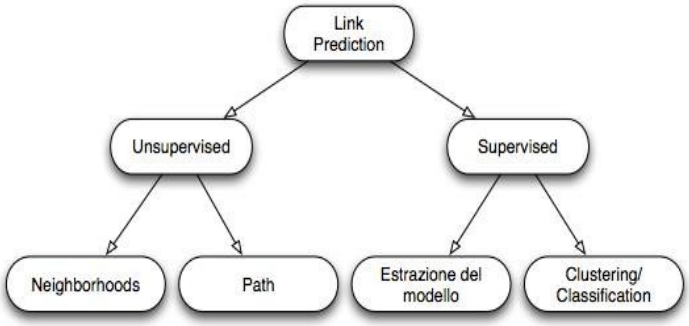

Figura 16: tipologie di metodi

2.1 Modelli Unsupervised

In questa tipologia, i metodi assegnano uno score a ciascuna coppia di nodi (x,y), usando come input le informazioni topologiche estratte dal grafo G: i risultati saranno quindi ordinati in una graduatoria in ordine di punteggio decrescente in modo da evidenziare gli archi la cui predizione è ritenuta più affidabile.

Questi metodi vengono adattati da tecniche utilizzate nella teoria dei grafi e nella social network analysis.
I modelli non supervisionati possono essere a loro volta suddivisi in due sotto-categorie: quelli basati su neighborhoods e quelli basati sull'analisi dei path [7,19].

Inoltre, sempre in questo modello, ne fanno parte i "meta-approcci", metodi che possono essere congiunti con le altre due sottocategorie[7].

Metodi basati sulla neighborhoods

Per ogni nodo x si analizza N(x), insieme che denota il set di vicini di x nella rete.
Numerosi approcci sono basati sull'idea che due nodi x e y abbiano più probabilità di formare un link nel futuro se la loro lista di vicini N(x) e N(y) ha un'ampia sovrapposizione; questo approccio segue l'idea intuitiva che, prendendo come esempio una rete di co-authorship, due autori che hanno molti colleghi in comune avranno maggiore probabilità di entrare in contatto fra loro.

Jin e Davidsen hanno definito modelli astratti per reti in evoluzione utilizzando questo principio, nel quale un arco (x,y) è più probabile che formi un arco (x,z) e uno (z,y) se nelle loro liste di vicini hanno entrambi z [7].

Common neighbors

L'implementazione più diretta è quella di definire lo score (x,y) secondo la formula:

$$score(x,y) = |\Gamma(x) \cap \Gamma(y)|$$

In questo modo vengono determinati il numero dei vicini che x e y hanno in comune.

Questa misura è stata utilizzata da Newman [7,20], il quale ha calcolato questa quantità nel contesto di reti di collaborazione, verificando una correlazione tra il numero di vicini comuni di x e y al tempo t e la probabilità che essi collaborino il futuro.

Jaccard

Il coefficiente di Jaccard, comunemente usato come misura di somiglianza nell'information retrieval [7], misura la probabilità che, selezionando una feature *f* appartenente ad x e y, questa sia presente sia in x sia in y.

Nel caso di Link Prediction, la feature si riferisce ai vicini nel grafo all'istante t.

Lo score viene assegnato tramite la formula:

$$score(x,y) = \frac{|\Gamma(x) \cap \Gamma(y)|}{|\Gamma(x) \cup \Gamma(y)|}$$

Questa misura è diversa da zero solo per le coppie di nodi per cui il valore di Common Neighbours è diverso da zero; a seconda dell'ordine degli score, i risultati dei due predittori coincidono abbastanza.

Adamic Adar

Questa misura di similarità nasce dall'idea di valutare la correlazione tra due pagine web in base ad una determinata feature. Nel caso di Link Prediction, la feature riguarda l'insieme dei vicini in comune ai due nodi.

La formula che permette la valutazione degli score è la seguente:

$$score(x,y) = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{\log |\Gamma(z)|}$$

Preferential attachment

L'osservazione alla base di questo modello è che la probabilità che un nuovo arco abbia il nodo x come suo estremo è proporzionale a |N(x)|, ossia al numero corrente dei vicini del nodo stesso.

Newman e Barabasi [7] hanno osservato che in reti di co-autori la probabilità che due nodi x e y siano collegati da un arco è correlata al numero di pubblicazioni associate ai due nodi.

La formula per calcolare lo score è la seguente:

$$score(x,y) = |\Gamma(x)| \times |\Gamma(y)|$$

Numerosi pubblicazioni hanno valutato la crescita delle reti [3,20,21,22].

Metodi basati sull'analisi dei path

Alcuni metodi sfruttano la distanza dei cammini minimi, considerando implicitamente l'insieme di tutti i percorsi tra due nodi. Vediamo alcuni metodi di analisi unsupervised che usano questa caratteristica.

Katz

La centralità di Katz è una misura che somma i pesi di tutti i path tra due nodi bilanciandoli sulla lunghezza tramite un fattore di decadimento esponenziale.

La formula è la seguente:

$$score(x,y) = \sum_{\ell=1}^{\infty} \beta^{\ell} \cdot |paths_{x,y}^{\langle \ell \rangle}|$$

dove paths è l'insieme della lunghezza dei cammini percorsi tra x e y, e
ß >0 è il parametro del predittore (per piccoli valori di ß gli score di predizione sono molto simili a quelli ottenuti con common neighbors, perché i percorsi di lunghezza 3 o più contribuiscono ben poco alla sommatoria).

Si può verificare che la matrice dei punteggi è data da (I-ß M)^-1 – I, dove M è la matrice di adiacenza del grafo.
Vengono considerate due varianti del coefficiente di Katz:

- non pesata, nella quale $|paths_{x,y}^{l}|$=1 se x e y hanno collaborato, e 0 se non hanno collaborato;

- pesata, nella quale $|paths_{x,y}^{l}|$ è il numero di volte che x e y hanno collaborato.

Hitting Time, PageRank e varianti

Un random walk su un grafo G inizia da un nodo x e si muove in modo iterativo da un vicino di x scelto uniformemente a caso dal set N(x).

Hitting Time ($H_{x,y}$) rappresenta il numero atteso di passi necessari di un random walk iniziato in x per raggiungere y. Poiché, Hitting Time non è una misura simmetrica, viene spesso considerata una sua variante il Commute Time, definita come: $C_{x,y}=H_{x,y}+H_{y,x}$.
Entrambe queste metriche rappresentano misure di prossimità e quindi possono essere usate come score per valutare la probabilità dell'esistenza di (x,y).

PageRank apporta delle modifiche al modello proposto da Hitting Time; in questo caso, durante il random walk, è introdotta la possibilità di avere dei reset casuali (random surfer). Viene definito uno score(x,y) parametrico rispetto ad **α**, il cui valore è compreso tra [0,1]. Esso definisce la probabilità in un cammino casuale di tornare ad x con un ad ogni passo.

SimRank

Lo score calcolato da SimRank è definito come il punto fisso della seguente definizione ricorsiva: due nodi sono simili nella misura in cui sono uniti a vicini simili.

Numericamente, questa quantità è specificata dalla seguente definizione di similarità:

$$\text{Similarity}(x,y) = \gamma \cdot \frac{\sum_{a \in \Gamma(x)} \sum_{b \in \Gamma(y)} \text{Similarity}(a,b)}{|\Gamma(x)| \cdot |\Gamma(y)|}$$

Il parametro γ è compreso tra 0 e 1 (dove similarity (x,x)=1).

Quindi lo score (x,y)= similarity (x,y).

"Meta-approcci"

Dopo aver presentato i modelli unsupervised, vengono introdotti due approcci di più alto livello che possono essere utilizzati al fine di ottimizzarne la fase analitica in modo da migliorarne le performance e semplificare la complessità di calcolo.

Gli approcci sono:

- low rank approximation;
- unseen bigrams.

Low-rank approximation

Sino ad ora abbiamo considerato come modello matematico di riferimento quello del grafo. Tale modello può essere espresso anche per mezzo di matrici di adiacenza: tutti gli algoritmi di Link Prediction, mostrati in precedenza, hanno una formulazione equivalente per tale modellazione.

La scelta di utilizzare matrici di adiacenza è spesso dettata da motivi, inerenti alla riduzione di complessità computazionale, che alcune operazioni su matrici consentono.
Nel caso d'uso di matrici di grandi dimensioni è tecnica comune quella di "ridurre" l'analisi su una matrice di rango k (con k sufficentemente piccolo) in modo da ridurre l'occupazione della matrice in memoria considerando solamente un sottoinsieme significativo della matrice di partenza. Questa riduzione può essere effettuata in modo efficiente tramite SVD (singular value decomposition).

Unseen bigrams

Il problema di Link Prediction è simile a quello di predire la frequenza di "coppie non presenti" nella modellazione di un linguaggio - coppie di parole che occorrono contemporaneamente all'interno di un testo, ma non nel testo utilizzato per effettuare il training.

Supponiamo di avere una la misura di score(x,y) già computata da uno dei predittori introdotti precedentemente: l'idea è quella di utilizzare tale score per predire score(y,z) per un nodo z simile ad x.

Viene definito Sx ^**δ** come l'insieme dei **δ** nodi più simili ad x secondo lo score (x, .), per **δ** appartenente a Z.
Lo score viene definito come:

$$\text{score}(x,y) = \left| \left\{ z : a \in \Gamma(y) \cap S_x^{\langle \delta \rangle} \right\} \right|$$

2.2 Modelli supervised

L'obiettivo di questa famiglia di modelli è quello di estrarre conoscenza dalla rete, al fine di poter effettuare previsioni. A differenza dei modelli unsupervised non vengono introdotte misure atte a calcolare gli score direttamente dall'utente ma, al contrario, le predizioni emergono da un processo di analisi che tende a estrarre conoscenza dalla stato osservato della topologia. Questi approcci garantiscono solitamente migliori performances a costo di una maggiore complessità metodologica.

In letteratura vi è un vasto studiato del problema di Link Prediction. Le strategie di Link Prediction possono essere classificate in quattro gruppi [33]:

- algoritmi di massima verosimiglianza;
- modelli supervised probabilistici;
- modelli supervised non probabilistici.

Il primo insieme di metodi è basato sulla stima di massima verosimiglianza. Studi empirici hanno dimostrato che molte reti reali hanno un'organizzazione gerarchica. Gli algoritmi che rientrano in questa categoria presuppongono alcuni principi di organizzazione della struttura della rete. Gli svantaggi di questo metodo sono:

- poca accuratezza;
- la richiesta di molto tempo per le analisi.

Il secondo gruppo di algoritmi è basato sul modello probabilistico Bayesiano. I modelli probabilistici mirano ad astrarre la struttura sottostante della rete osservata e prevedere gli archi mancanti utilizzando il modello appreso. Data una rete, il modello probabilistico ottimizza una funzione target al fine di stabilire quale sia il modello che meglio descrive il fenomeno osservato. Quindi, la probabilità che con cui un nuovo link apparirà nella rete è stimata tramite una probabilità condizionata calcolata

sullo specifico modello identificato.

L'ultimo gruppo di metodi proposti utilizza approcci supervisionati (introdotto in letteratura in [7]).
Un primo esempio di questo metodo è lo studio sull'utilità delle feature topologiche di una rete di co-autori. In letteratura sono state proposte numerose caratteristiche di similarità per coppie di nodi: per aumentare la precisione delle predizioni, su di esse viene quindi addestrato e utilizzato un classificatore. Esso avrà il compito di prevedere i collegamenti futuri tra nodi.
Recentemente sono stati proposti nuovi modelli che sfruttano un approccio supervisionato utilizzando strategie molto diverse fra di loro. In [34] sono usate le caratteristiche topologiche e gli attributi dei nodi di una rete sociale dinamica, applicando una combinazione lineare ad una matrice di covarianza, per ottimizzare la previsione. La componente principale di regressione è l'algoritmo utilizzato in [35] per determinare il peso delle variabili predittive, statisticamente indipendenti, utilizzate per la previsione.
Un altro approccio è quello di Rank Aggregation, proposto in [36] in cui vengono classificati i nodi secondo alcune misure topologiche; al nuovo istante di tempo ogni misura è ponderata in base alla performance di previsione. In [37], su una rete di pubblicazioni, sono usate caratteristiche testuali e topologiche al fine di prevedere nuove citazioni, applicando un metodo di apprendimento supervisionato tramite l'adozione di un classificatore SVM. In [45] è usata una procedura di fattorizzazione di tensori per selezionare gli attributi più predittivi, mentre in [38] importanti variabili per il link prediction sono esaminate ed è fornito forniscono un framework predittivo avente alte.

I metodi supervisionati si sono spesso dimostrati migliori sia in termini di prestazioni sia in termini di accuratezza rispetto ai modelli non supervisionati. Dato l'interesse per le grandi reti sparse, è importante concentrarsi sulle informazioni locali raccolte dalle comunità e sulle caratteristiche temporali per addestrare un classificatore, al fine di

prevedere interazioni future.
Diversi approcci, come in [39], usano la tecnica del clustering e dell'informazione delle comunità per migliorare i metodi di link prediction. Le analisi fornite dimostrano che utilizzare l'output di approcci di clustering su reti migliora l'accuratezza della previsione di nuovi collegamenti.
Altri lavori [46,47] mostrano come, analizzando reti dinamiche, un approccio basato sul'analisi di serie temporali, che hanno modellino la continua evoluzione delle features inerenti le connessioni tra i nodi della rete, possa contribuire nella soluzione del compito predittivo.

Al fine di costruire un classificatore efficace per il link prediction, è fondamentale definire e calcolare una serie di caratteristiche strutturali del grafo. Quando si tratta reti di grandi dimensioni, le quali possono includere milioni di vertici e archi, una delle sfide è quella dell'estrazione computazionale di tali caratteristiche. Diversi studi relativi al link prediction [40,41,42,43,44] dimostrano che utilizzare caratteristiche appropriate è cruciale ai fini della predizione.
Ad esempio, in [40], si analizza la relazione tra la struttura della rete e la performance dell'algoritmo di predizione, mentre in [41] è mostrato che solo piccoli insiemi di caratteristiche sono essenziali per prevedere nuovi archi; inoltre i nodi con alta centralità sono più predittivi rispetto ai nodi con bassa centralità. Infine, da questi lavori, è emerso che in reti con basso coefficiente di clustering, i metodi di previsione hanno scarso rendimento, mentre, come il coefficiente di clustering cresce, l'accuratezza migliora drasticamente.

In conclusione, si può affermare che i metodi supervised si possono suddividere in due sottoclassi di algoritmi [16, 19,24, 25, 26, 27]:

- quelli che estraggono il modello evolutivo della topologia del grafo per sfruttarne le informazioni per poi effettuare le predizioni;
- quelli che affiancano al Link Prediction tecniche di data mining, generalmente classificazione e clustering, in modo da poter

sfruttare, oltre le informazioni topologiche della rete, anche quelle contestuali.

Estrazione del modello dalla rete

In questa tecnica le performance che ottengono i predittori sono legate al grado in cui il modello rispecchia quello delle singole reti.

Una strategia spesso adottata in questo contesto applicativo è quella di sfruttare le informazioni topologiche della rete e per cercare di produrre delle regole capaci di identificarne il modello evolutivo e, in seguito, essere utilizzarle per guidare il compito di Link Prediction.

Due algoritmi che seguono questo principio sono Germ (Graph Evolution Rule Miner) e la sua variante LFR [28,29]. Essi permettono lo studio evolutivo della rete partendo dalle regole che estratte dalla rete stessa.

In figura 17 una schematizzazione dell'algoritmo GERM: a sinistra sono riportati esempi di regole estratte (composte da un body – stato osservato - e un head – ultimo arco apparso) e a destra una loro applicazione (l'arco rosso è quello predetto).

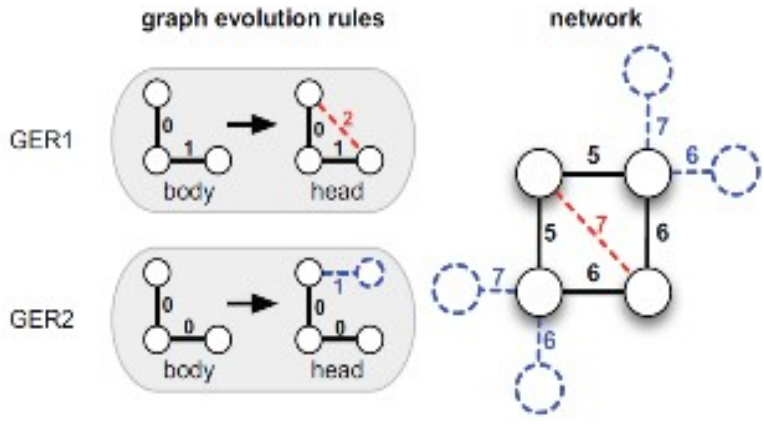

Figura 17: l'algoritmo GERM

Classificazione e clustering

Come già discusso, in molti lavori di ricerca [30,31] è emerso che affiancare in maniera iterativa gli algoritmi di Link Prediction, di tipo unsupervised, ad algoritmi di classificazione o clustering permette di migliorare in maniera consistente le performances finali rispetto alla singola adozione degli stessi.

Molte reti presentano un pattern che è definito Community Structure [1,3].
Una community è considerata come un insieme di nodi in cui ogni nodo è più vicino ad altri nodi all'interno di una comunità rispetto ai nodi esterni ad essa. Tale fenomeno è presente in molte reti reali, soprattutto in contesti sociali.
Una volta raccolte informazioni topologiche relative alle community presenti nel grafo queste possono essere utilizzate all'interno del processo predittivo: come vedremo nella sezione sperimentale di questa tesi, fornire tali informazioni ad un classificatore può riflettersi in un valido incremento dell'accuratezza delle predizioni.

2.3 Confronto fra i due modelli

Considerando i modelli unsupervised possiamo notare come le informazioni sulla topologia della rete siano estratte in modo diretto e mirato soggetto ad una scelta esplicita del'utente. L'utente decide a priori quale caratteristica della rete meglio si presta a spiegarne l'evoluzione: la bontà delle predizioni è quindi strettamente connessa alla correttezza dell'assunzione effettuata in partenza.
Un grosso limite, dovuto a tale imposizione, è identificabile nel fatto che i migliori predittori unsupervised (Adamic Adar e Katz in contesti sociali) hanno accuratezza molto bassa, compresa tra il 10% e il 16%.

Rispetto ai modelli non supervisionati l'adozione di un approccio supervisionato riesce a garantire migliori performances. Ovviamente dalla letteratura emergono altri limiti quali, ad esempio, una maggiore complessità di design ed un minore controllo del processo decisionale [32].

Possiamo quindi concludere che prevedere l'evoluzione di una rete non è un compito facile, poiché:

- le reti sono contenitori di legami deboli;
- molte previsioni risultano essere falsi positivi;
- gli approcci semplici non sono buoni;
- gli approcci complessi hanno costi computazionali elevati.

PARTE III
I CLASSIFICATORI

Dopo aver dato la definizione di classificazione nella sezione Stato dell'arte (4.1), introduciamo le tecniche e gli algortimi di classificazione che saranno in seguito utilizzati nella sezione sperimentale al fine di prevedere i nuovi archi.

Le tecniche di classificazione discusse appartengono a tre famiglie, in particolare discuteremo di :

- Alberi di decisione (Decision Tree);
- Support Vector Machine (SVM);
- Classificatori basati sul teorema di Bayes.

1. Alberi di decisione

Gli alberi di decisione sono la tecnica di classificazione maggiormente utilizzata in letteratura: un decision tree permette di rappresentare con un albero un insieme di regole di classificazione.

Prima di spiegare il funzionamento di un albero di decisione, è necessario descriverne la struttura.
Un albero di decisone è composto da:

- i nodi interni, i quali sono attributi del soggetto da classificare;
- gli archi in uscita da un nodi etichettati con i valori che l'attributo può assumere;
- i nodi foglia, rappresentanti le classi.

L'idea alla base degli alberi di decisione è che il procedimento di classificazione possa essere modellato tramite una sequenza di semplici quesiti [48,49].
Ogni quesito della sequenza può essere visto come un nodo. Il primo nodo è detto radice: da esso ha inizio il procedimento di costruzione dell'albero. Ad esso sarà associata una feature di split e un valore di soglia usato per suddividere i dati in input.

La costruzione del nodo radice porta quindi alla divisione dell'insieme di dati in due sottoinsiemi disgiunti di oggetti: il sottoinsieme per i cui elementi la condizione nella radice è vera e quello per i cui elementi essa è falsa. Al termine di ciascun ramo può esserci un nuovo nodo o una foglia, cioè un nodo terminale che non pone alcun quesito, ma contiene un'etichetta di classe.
Questo tipo di albero è detto monotetico, ovvero ciascun nodo verifica una sola caratteristica dell'oggetto, e binario, perché ogni nodo ammette non più di due risposte.

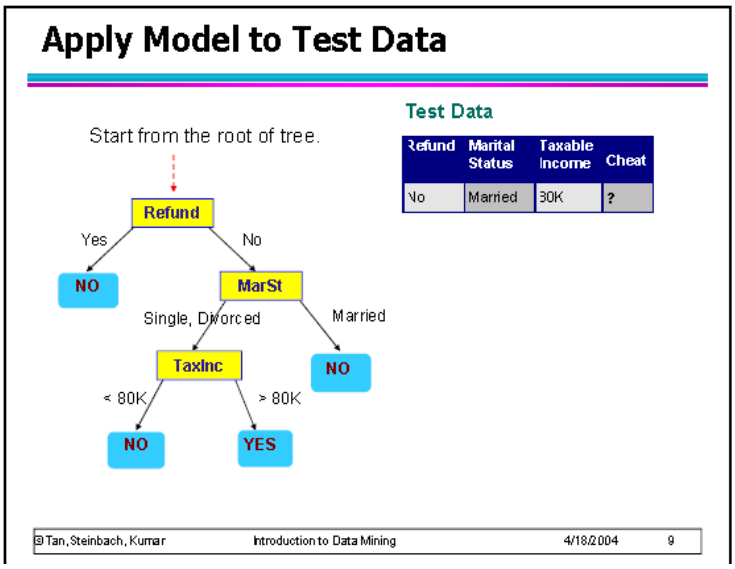

Figura 18: esempio di alberi di decisione

Il classificatore percorre quindi il ramo relativo all'esito della valutazione e visita il nodo ad esso collegato, detto nodo figlio.

Il processo di costruzione si ferma quando viene soddisfatto un criterio d'arresto, ad esempio il criterio di perfezione: quando ad un nodo arrivano solo oggetti della stessa classe, la loro etichetta è attribuita al nodo, il quale diventa una foglia.
Un albero che classifica correttamente tutti gli elementi dell'insieme dei dati è detto perfetto o completo. Inoltre, se il numero di decisioni da compiere è lo stesso per ciascuna foglia l'albero è detto bilanciato. In figura 1 viene riportato un esempio di albero decisione.

La costruzione di un albero si articola nelle seguenti fasi [2]:

- Selezione delle variabili di splitting;
- Definizione del numero di ramificazioni;
- Valutazione della miglior suddivisione e tipologia di partizione;

- Termine della procedura di crescita;

- Potatura.

Selezione delle variabili di splitting

La suddivisione del dataset ha lo scopo di individuare le variabili che spiegano maggiormente le osservazioni. Una tecnica utile a tale fine è quella di mettere le variabili più importanti nei nodi più vicini alla radice. Con questo passo, ad ogni nodo, l'algoritmo cerca quale attributo produce la suddivisione migliore rispetto alle osservazioni disponibili rispetto all'insieme dei dati che verificano le condizioni sui nodi precedenti.

Numero di ramificazioni

Ogni nodo padre può generare due o più nodi figli.

La maggior parte degli algoritmi di classificazione opera una suddivisione binaria in modo da non ridurre il numero di osservazioni in ciascun nodo figlio; ovviamente la suddivisione in più nodi figli può essere riprodotta mediante una sequenza di split binari.

Valutazione della migliore suddivisione e tipologia di partizione

L'individualizzazione dell'attributo e del suo relativo valore soglia che comporta la miglior suddivisione è basata sulla massimizzazione di qualche misura di separazione di osservazioni del nodo corrente.

La maggior parte degli algoritmi produce una partizione ricorsiva; ciascuna operazione di suddivisione viene eseguita massimizzando una funzione di interesse locale, senza badare alle suddivisioni futuri. Ciò avviene poiché identificare la suddivisione migliore comporta un costo computazionale molto alto dal momento che si dovrebbero generare tutti i possibili sotto-alberi seguenti prima di confermare ciascuna partizione.

Termine della procedura di crescita

La crescita, cioè la successione delle operazioni di suddivisione, deve terminare prima di conformare l'albero rispetto ai dati di stima.

Generalmente sono usate regole basate sul numero minimo di foglie, sul numero minimo di osservazioni in ciascuna foglia e sulla profondità dell'albero per decidere quando arrestare la procedura generativa.

Potatura

Tale fase consiste nel ripercorrere l'albero generato a ritroso al fine di ottimizzare la sua dimensione in base a diversi criteri, quali:

- l'utilizzo di un insieme di dati di verifica su cui testare il modello;
- l'utilizzo di una funzione costo-complessità associata all'algortimo.

Misure di impurità

La costruzione di un albero porta alla suddivisione dell'insieme di dati a disposizione in regioni organizzate gerarchicamente. Una regione è detta pura se contiene solo oggetti appartenenti alla medesima classe. L'impurezza di un nodo può essere stimata con diversi metodi:

- Indice di Gini:

$$GINI(i) = 1 - \sum_{j=1}^{k} [p(j \mid i)]^2$$

dove p(j|i) è la frequenza della classe j rispetto al nodo i;

- Entropia:

$$Entropy(i) = -\sum_{j=1}^{k} p(j \mid i) \log p(j \mid i)$$

dove p(j|i) è la frequenza della classe j rispetto al nodo i;

- Errore di classificazione:

$$Error(i) = 1 - \max_{j \in K} p(j \mid i)$$

L'indice di Gini e l'Entropia sono di solito utilizzati per guidare la costruzione dell'albero, mentre la valutazione del tasso di errore nella classificazione è utilizzato per effettuare un'ottimizzazione dell'albero, nota come processo di pruning o potatura. Poiché, in generale, in un buon albero di decisione i nodi foglia sono il più possibile puri, un'ottimizzazione dell'albero consiste nel cercare di minimizzare il livello di entropia man mano che si scende dalla radice verso le foglie. In tal senso, la valutazione dell'entropia determina quali sono, fra le varie scelte a disposizione, le condizioni di split ottimali per l'albero di classificazione.

Gli alberi di decisione hanno le seguenti caratteristiche:

- accuratezza della previsione;
- velocità: tempo impiegato sia per costruire il modello sia per classificare gli elementi;
- scalabilità;
- robustezza: capacità del modello di classificare correttamente elementi anche in presenza di dati erranti o mancanti;
- interpretabilità: il modello che viene creato è di facile comprensione.

1.1 Tipologie di algoritmi decision tree

Gli algoritmi trees che sono stati usati nella sezione sperimentale e che verranno analizzati in 1.1.1, in 1.1.2 e in 1.1.3 sono:

- C 4.5 (in weka chiamato J48);
- Bagging;
- Random Forest.

1.1.1 C4.5

C 4.5 è il principale algoritmo utilizzato per generare un albero di decisioni, sviluppato da Ross Quinlan [53,54]. L'idea che sta alla base dell'algoritmo è quella di costruire l'albero di decisione attraverso il frazionamento del dataset in sottoinsiemi sempre più piccoli, utilizzando l'entropia come criterio per lo split.

La costruzione dell'albero si interrompe quando:

- il nodo contiene record appartenenti a una sola classe;
- il nodo non contiene record.

Il Bagging e Random Forest sono evoluzioni di tale algoritmo.

1.1.2 Bagging

L'algoritmo Bagging provvede a campionare ripetutamente il training set di input e, differentemente da altri approcci supervisionati, prevede il reinserimento dei record già selezionati durante la fase di costruzione. Ciò avviene ciclicamente su un diverso sample del training set, a partire da un campione di bootstrap, al fine di migliorare progressivamente le performance del modello decisionale.

Un campione di bootstrap è un campione casuale i cui elementi vengono estratti con reinserimento, per questo la tecnica prende anche il nome di bootstrap aggregating. Tutti gli insiemi costruiti ad ogni iterazione sono di taglia uguale a quella del training set di partenza. Poiché si campiona con reinserimento è possibile che in uno stesso insieme siano presenti elementi ripetuti e che elementi del training set vengano omessi [49,50,51,52].

Questa tecnica permette di stabilizzare i risultati ottenibili del data set a disposizione. È utile nel caso di data set di piccole dimensioni.

1.1.3 Random Forest

Le foreste casuali sono una tecnica di classificazione, sviluppate da Breiman e Culler [55,56], utilizzate per migliorare le prestazioni di un processo di analisi.
L'idea che sta alla base delle foreste casuali consiste nelle combinazione di alberi ottenuti utilizzando la selezione casuale delle variabili esplicative.

L'algoritmo prende in input un sottoinsieme di features, scelte casualmente tra quelle disponibili, e le combina tra di loro. Ripete tale operazione, finché non termina il numero di combinazioni a disposizione. A questo punto, le confronta e sceglie l'albero migliore.
In figura 19, lo schema delle Random Forest.

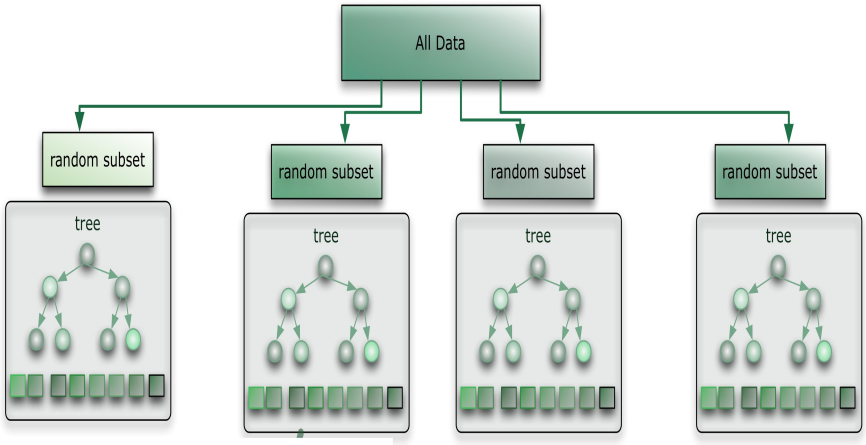

Figura 19: schema delle Random Forest

I vantaggi di questo approccio sono:

- Produzione di un albero classificatore molto accurato con un insieme di dati molto grande, riuscendo a dare una stima di importanza alle variabili esplicative;

- Gestione di dataset aventi un gran numero di features e identificazione delle loro interazioni;

- Non sensibile a record costituiti da dati parziali o mancanti;

- Utile per l'individuazione di outliers e per la visualizzazione dei dati mancanti.

2. Support Vector Machine (SVM)

Le Support Vector Machine (SVM), sviluppate negli anni '90 da Vladimir Vapnik [57,58], sono un approccio alla classificazione e alla regressione, basato su due fasi:

- sfruttare l'informazione rappresentata dai punti situati "all'interfaccia" fra le classi;

- trasformazione dello spazio delle features per esprimere in termini lineari un problema di classificazione non lineare.

Di solito i dati sono portati dallo spazio originale (per il quale non esiste un semplice criterio di separazione delle classi target) allo spazio trasformato di più facile interpretazione (tramite una trasformazione non lineare, figura 20). Una volta eseguita tale trasformazione, la separazione dei parametri può essere eseguita individuando una retta, un piano o un iperpiano, a seconda della dimensione dello spazio dei parametri, scegliendo poi quello con il margine massimo.

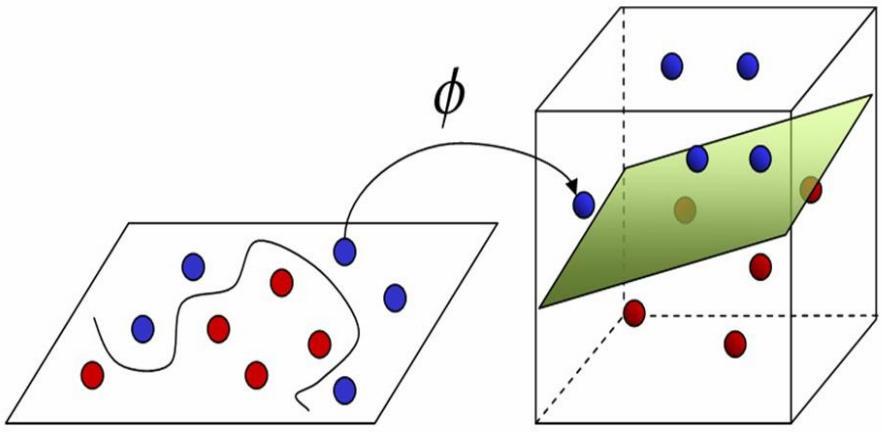

Figura 20: Trasformazione dallo spazio originale allo spazio trasformato e individuazione dell'iperpiano.

Supponendo *margine* la somma delle distanze dell'iperpiano separatore dai campioni di training set ad esso più vicini di ciascuna classe (per ciascuna classe si considera la distanza una sola volta, anche se il campione più vicino non è unico).
Questo metodo ha due vantaggi:

- Può generare un modello non lineare di classificazione;

- Previene l'overfitting (finché il criterio di decisione è lineare nello spazio trasformato).

Esiste una variante al metodo SVM, chiamata SVM- soft, che permette di massimizzare il margine e contemporaneamente minimizzare gli errori sul data set.

Estendendo il problema all'utilizzo di SVM come classificatore non lineare, l'idea di base consiste nel trasformare non linearmente lo spazio delle features e nel calcolare poi i prodotti scalari nello spazio trasformato. Per prodotto scalare si intende un'operazione binaria che associa ad ogni coppia di vettori, appartenenti ad uno spazio vettoriale definito sul campo reale, un elemento del campo.

L'approccio SVM ha, quindi, proprietà importanti sul piano della capacità di generalizzazione. E' possibile dimostrare che l'addestramento di una SVM equivale alla minimizzazione dell'errore empirico di classificazione e, allo stesso tempo, alla massimizzazione del margine geometrico.

Alcune applicazioni per cui le SVM sono state usate con successo sono:

- Elaborazione del linguaggio naturale;

- Classificazione di testi;

- Ricerca bioinformatica.

Il limite di questo algoritmo è la complessità computazionale.

3. Classificatori bayesiani

I classificatori bayesiani rappresentano un approccio probabilistico alla risoluzione di problemi di classificazione; tali oggetti modellano relazioni probabilistiche tra gli attributi descrittivi (features) e l'attributo di classificazione (target).

Questa famiglia di classificatori trova il suo fondamento teorico sull'applicazione del teorema di Bayes [50,59].

3.1 Teorema di Bayes

Il teorema di Bayes deriva da tre teoremi fondamentali delle probabilità: il teorema della probabilità condizionata ed il teorema della probabilità composta.

In teoria della probabilità, la probabilità condizionata di un evento A rispetto a un evento B identifica la probabilità che si verifichi A , sapendo che B è verificato. Più formalmente:

$$P(A|B) = P_B(A) = \frac{P(A \cap B)}{P(B)}$$

Il teorema della probabilità composta deriva da quello della condizionata:

$$\mathrm{P}(A \cap B) = \mathrm{P}(B)\mathrm{P}(A|B) = \mathrm{P}(A)\mathrm{P}(B|A),$$

per cui la probabilità che due eventi si verifichino contemporaneamente è pari alla probabilità di uno dei due eventi moltiplicato con la probabilità dell'altro evento condizionato dal verificarsi del primo.

Il teorema di Bayes nella sua forma più semplice è scritto nel seguente modo:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(B|A)\,P(A)}{P(B)}.$$

Gli eventi A e B possono essere eventi qualunque sia contemporanei, sia aventi luogo in tempi diversi. Esso viene impiegato per calcolare la probabilità di una causa che ha scatenato l'evento verificato.

Il teorema di Bayes costituisce il fondamento dell'inferenza bayesiana.
Per inferenza bayesiana si intende un approccio all'inferenza statistica (procedimento per cui si inducono le caratteristiche di una popolazione dall'osservazione di una parte di essa, detta campione, selezionata solitamente mediante un esperimento casuale) in cui le probabilità non sono interpretate come frequenze, proporzioni o concetti analoghi, ma piuttosto come livelli di fiducia nel verificarsi di un dato evento.

Dalle definizioni del teorema di Bayes e di inferenza bayesiana si sviluppa l'approccio bayesiano.
Esso si pone come obiettivo il problema di fare delle previsioni e ritiene il problema della formulazione di ipotesi a partire dai dati, come un suo sotto-problema. Un modo per specificare che cosa si intende per la migliore ipotesi è quello di affermare che la migliore ipotesi è quella più probabile, avendo a disposizione dei dati ed una certa conoscenza iniziale delle probabilità a priori delle varie ipotesi. Le ipotesi elaborate dai dati e combinate in modo opportuno, portano alla formulazione di una previsione.

3.2 Le reti Bayesiane

Un altro concetto importante quando si tratta di classificatori bayesiani è quello di rete bayesiana [60].

Una rete bayesiana è la rappresentazione grafica di un modello probabilistico, cioè la riproduzione di una distribuzione di probabilità su un insieme di variabili, tramite un grafo aciclico diretto.
Un grafo aciclico diretto (DAG) è un particolare tipo di grafo diretto che non presenta cicli diretti, ovvero comunque sia scelto un vertice del grafo non è possibile ritornare ad esso percorrendo gli archi del grafo. Un grafo diretto può dirsi aciclico (cioè è un DAG) se una visita in profondità non presenta archi all'indietro (figura 21).

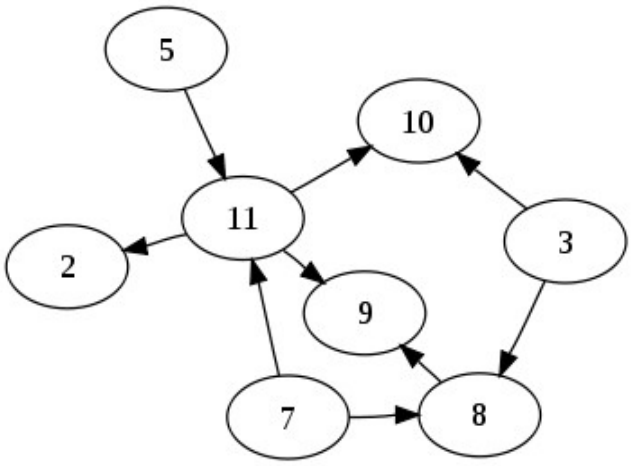

Figura 21: esempio di grafico aciclico diretto

I grafi aciclici diretti sono composti da:

- nodi, che rappresentano le variabili;

- archi, che rappresentano le relazioni di dipendenza statistica tra le variabili e le distribuzioni locali di probabilità dei nodi foglia rispetto ai valori dei nodi padre.

Una rete bayesiana, quindi, rappresenta la distribuzione della probabilità congiunta di un insieme di variabili. In figura 22 viene riportato un esempio di rete bayesiana.

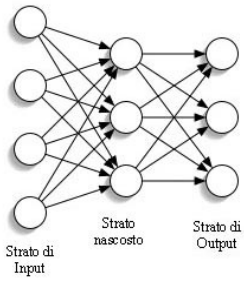

Figura 22: esempio di rete bayesiana

Le caratteristiche di una rete bayesiana sono:

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

- il grafo non ha cicli diretti.

I vantaggi dell'utilizzo delle reti bayesiane sono:

- la rappresentazione grafica e la struttura delle relazioni tra variabili aleatorie risulta intuitiva e di facile comprensione;

- utilizzabili per insiemi di dati incompleti.

3.3 Tipologia di algoritmi bayesiani

Il classificatore bayesiano richiede la conoscenza delle probabilità a priori e condizionali relative al problema, quantità che in generale non sono note ma sono tipicamente stimabili. Se è possibile ottenere delle stime affidabili delle probabilità coinvolte nel teorema, il classificatore bayesiano risulta generalmente affidabile e potenzialmente compatto.
E' molto usato nella classificazione dei testi e in medicina.
Gli algoritmi bayesiani che sono stati usati nella sezione sperimentale e che

verranno di seguito analizzati in sono:

- Naive Bayes;
- Bayes Net.

Naive Bayes

Il classificatore Naive Bayes [49,59,60] è basato sul concetto che tutti gli attributi che descrivono una certa istanza di classificazione sono tra loro condizionatamente indipendenti data la categoria a cui appartiene l'istanza.

L'assunzione di indipendenza delle features di classificazione permette di apprendere separatamente i parametri di ogni attributo, semplificando molto l'apprendimento specialmente in quelle situazioni in cui il numero di attributi è elevato ed in cui i dati a disposizione non sono molto numerosi.

Il dominio applicativo del classificatore Naive Bayes riguarda la classificazione di istanze che possono essere descritte mediante un insieme di attributi di cardinalità elevata. Esso opera stimando i parametri del modello utilizzando i dati di training a disposizione: tale stima corrisponde all'identificazione di uno specifico un modello tra tutti quelli presenti nello spazio delle ipotesi.
Il modello prescelto sarà quindi applicato per classificare le nuove istanze.

A differenza di altri algoritmi di apprendimento, non c'è un'esplicita ricerca nello spazio delle possibili ipotesi, ma l'ipotesi viene definita semplicemente contando la frequenza degli attributi negli esempi di addestramento. Il metodo di apprendimento bayesiano si è rivelato utile in molti contesti applicativi tra cui la classificazione di documenti testuali.

Bayes Net

Alla base del classificatore Bayes Net troviamo il concetto, precedentemente introdotto, di reti bayesiane.

I nodi rappresentano variabili casuali in senso Bayesiano: possono essere quantità osservabili, variabili latenti, parametri sconosciuti o ipotesi. Gli archi rappresentano condizioni di dipendenza; i nodi che non sono connessi rappresentano variabili che sono condizionalmente indipendenti tra di loro. Ad ogni nodo è associata una funzione di probabilità che prende in input un particolare insieme di valori per le variabili del nodo genitore e restituisce la probabilità della variabile rappresentata dal nodo.

Come Naive Bayes, il principale contesto applicativo per Bayes Net è quello relativo alla classificazione di documenti testuali.

PARTE IV
SEZIONE SPERIMENTALE

1. Introduzione

Le reti complesse sono oggi utilizzate per descrivere una vasta gamma di scenari del mondo reale: interazioni sociali e biologiche così come sistemi economici sono esempi di come ampia sta diventando la varietà di argomenti che vengono studiati tramite la scienza delle reti.

Alcuni problemi di rete sono stati analizzati: la scoperta di comunità, il link prediction, e la classificazione sono alcune delle diverse attività studiate.

Tra tutti questi compiti, i più interessanti sono finalizzati a descrivere come le reti si evolvono nel tempo.

Le reti sono raramente utilizzate per modellare entità statiche: se si considera, ad esempio, una rete sociale, si può notare come, con il passare del tempo, nuovi utenti appaiono o scompaiono, avvengono nuove interazioni o si interrompono.

La comprensione di queste dinamiche è il primo passo per ottenere intuizioni sulla vera natura del fenomeno modellato dalle reti. Inoltre quasi tutti i problemi di rete possono essere riformulati tenendo conto della dimensione temporale: le comunità possono essere monitorate e tramite lo studio del loro ciclo di vita è possibile ripercorrerne la storia; nuovi link possono essere previsti utilizzando le informazioni sull'analisi dei nodi.

Le reti che tengono conto della dimensione temporale sono chiamate dinamiche. La topologia di queste reti si evolve nel corso del tempo e nuovi nodi e nuovi archi possono apparire a seconda delle interazioni fra utenti. La previsione della probabilità di una futura associazione tra due nodi, sapendo che non c'è alcuna associazione tra i nodi nello stato attuale del grafo, viene comunemente detto link prediction.

In generale, il link prediction fornisce una misura di prossimità sociale tra due vertici in un gruppo, che se noto, può essere utilizzata per ottimizzare gli obiettivi sull'intero gruppo.

Rivolgendosi all'approccio del link prediction tramite l'apprendimento supervisionato, a seconda delle reti analizzate, alcune misure sono risultate interessanti. Sono state considerate alcune caratteristiche delle reti sociali come la struttura di comunità, le misure di centralità, di rete, di node ranking e quelle relative alla link prediction.

Per ogni rete vengono calcolate un insieme di features che descrivono le

caratteristiche per ogni coppia di nodi.

Ogni valore di ogni caratteristica viene utilizzato per costruire il classificatore per prevedere nuovi collegamenti.

Sono state utilizzati una serie di algoritmi di classificazione per valutare le loro performance sul problema e successivamente viene fatta un'analisi comparativa tra questi, utilizzando diversi parametri di rendimento (Accuratezza, AUROC).

Infine, si dimostra che le performances sono risultate buone.

2. Dataset utilizzati

In questo capitolo vengono descritte e analizzate le tre reti utilizzate.

2.1 Rete Foursquare-Osaka-1h-RealNet

La rete di Foursquare è costituita su un campione di utenti di Osaka (Giappone).

Le informazioni di base di tale rete sono le seguenti:

- rete non diretta;
- |V|= 2642;
- |E|= 24164;
- fonte: www.foursquare.com

Prima di addentrarci nell'analisi vera e propria della rete di Foursquare, è bene dare alcune informazioni su tale applicazione basata sulla geolocalizzazione.

Lanciata nel marzo 2009, Foursquare è un'applicazione mobile e web che permette agli utenti registrati di condividere la propria posizione con i propri contatti. Gli utenti eseguono il check-in, cioè la registrazione della propria posizione, tramite la versione browser del sito o tramite applicazioni su dispositivi che utilizzano il GPS. La registrazione della propria posizione permette la retribuzione o con punti che permettono di scalare una classifica settimanale nella quale fanno parte i contatti della stessa città, o con il ricevimento di badge, ossia il riconoscimento per aver raggiunto certi obiettivi, eseguendo il check-in in certi luoghi, ad una certa frequenza o trovandosi in una certa categoria di luoghi.

I check-in possono essere condivisi, insieme ad un breve status, collegando Foursquare ai propri profili Facebook e Twitter.

Infine, Foursquare permette agli utenti di:

creare badge personalizzati;

creare liste pubbliche accompagnate da brevi suggerimenti per gli utenti che eseguono la registrazione nel luogo stesso.

2.1.1 Analisi

Con l'utilizzo del software Cytoscape (www.cytoscape.org), si è potuto procedere all'analisi di seguito riportata. Per prima cosa è stata evidenziata la struttura della rete di Foursquare, come si riporta in figura 23.

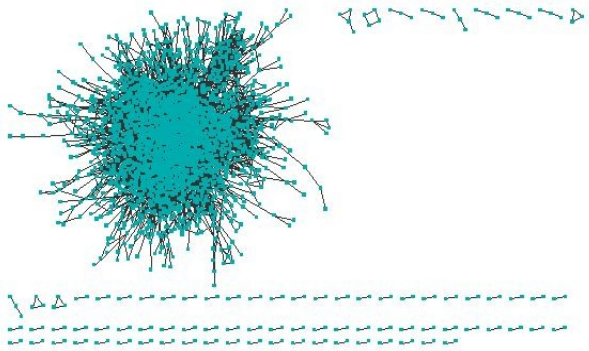

Figura 23: Struttura della rete Foursquare

Si nota la presenza di una componente gigante (dove per componente gigante si intende una componente connessa che contiene una frazione significativa di tutti i nodi) e, vicino a questa, delle altre piccole componenti connesse che sono formate da un minimo di due nodi e un arco che li unisce a un massimo di quattro nodi e quattro archi totali.

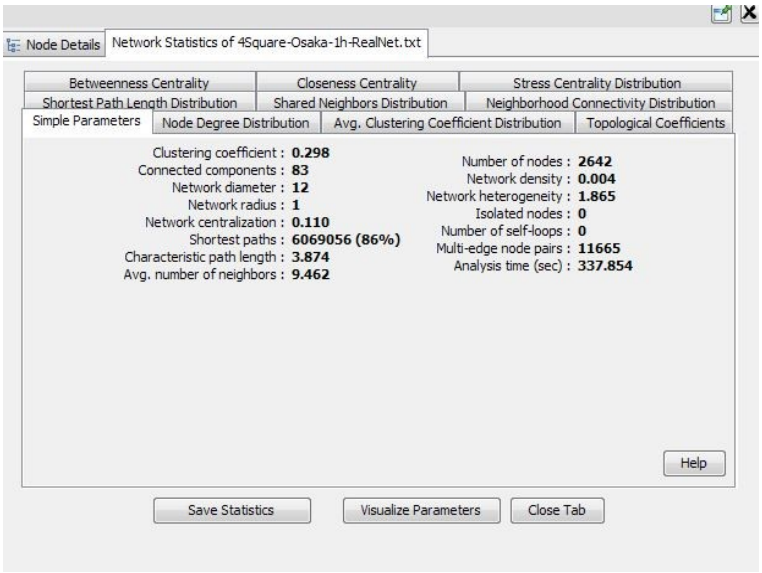

Figura 24: i dati della rete

Partendo dall'analisi delle misure di rete (figura 24), la prima che andiamo a analizzare è il coefficiente di clustering. Questo parametro stima quanto i nodi adiacenti ad un altro nodo siano anch'essi in relazione fra loro; esso può assumere valori compresi nell'intervallo tra [0,1]; in questo caso, il valore del coefficiente di clustering è di 0,298, quindi ci sono poche utenti che sono registrati nello stesso luogo.

La seconda misura analizzata riguarda la frammentazione della rete. In questo caso si hanno 83 componenti connesse, come è possibile notare in figura 1.

La terza misura riguarda il diametro della rete, ossia la massima distanza fra ogni coppia di nodi del grafo. In questo caso assume un valore pari a 12.

Il *network radius* rappresenta il minimo valore tra le eccentricità dei nodi. Il raggio della rete è 1.

Il parametro della *Network centralization* ha un valore di 0,110. Lo studio della centralità svolge un ruolo importante perché permette di: giudicare la rilevanza o la criticità dei nodi della rete,attribuire una misura fra i nodi o le aree di una rete, identificare il grado di coesione di un'area della rete e, infine, identificare le aree di una rete. Nello studio delle reti complesse si hanno tre nozioni di centralità: *Degree centrality, Closeness Centrality, Betweenness Centrality,* le quali verranno analizzate successivamente. Le tre misure di centralità identificano anche tre definizioni diverse di aree della rete. L'idea di base è che un'area sia una zona nella quale i nodi abbiano una coesione tra loro: coesione in base al numero di connessioni dirette; coesione in base al numero di connessioni a una certa distanza; coesione in base alla resistenza alla rimozione di nodi. In questo caso il valore è basso, molto probabilmente perché i nodi non fanno tutti capo a un unico centro.

La misura relativa allo *shortest path,* ossia ai cammini più brevi, ha un valore pari a 6069056 (86%); questo significa che i cammini esistenti (6069056) sono l'86% dei cammini possibili.

La misura riguardante la *characteristic path lenght,* ossia la lunghezza dei cammini, ha un valore pari a 3,874. Confrontando questo valore con il grafico relativo alla *shortest path lenght distribution* (Figura 25), si nota che la moda si aggira intorno a 4.

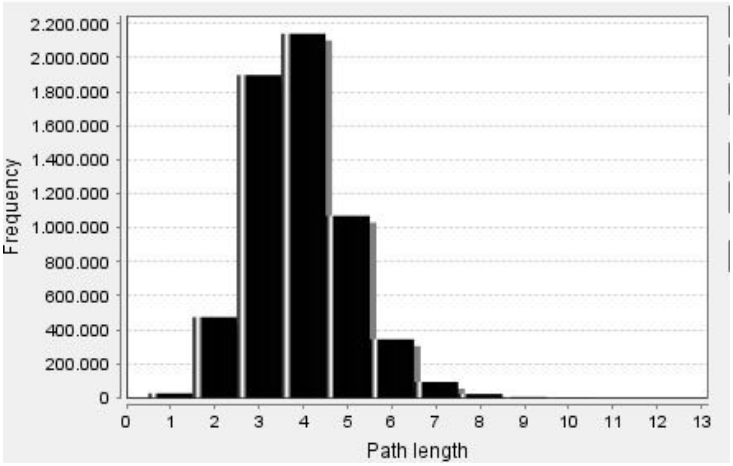

Figura 25: shortest path distribution rete foursquare

La misura relativa all'*Avg. Number of neighbors* ha un valore pari a 9.462, ovvero ogni nodo ha attorno a sé circa quel numero di vicini.

La densità identifica quanti archi sono presenti nella rete rispetto ai possibili archi. Il suo valore è pari a 0.004.

Una caratteristica della rete è che non sono presenti nodi isolati, in quanto ciascun nodo presente può essere messo in relazione con qualunque altro nodo grazie al fatto che tutti i nodi possono essere connessi tra loro attraverso i vari cammini.

Il *number of self loops* ha valore 0; ciò è dovuto al fatto che la rete sociale in questione non permette, infatti, di creare relazioni di amicizia con se stessi.

Analizzando la media del coefficiente di clustering,con gli assi impostati su scala lineare, si nota (figura 26) che con un numero di vicini basso (27) si

ha una media di coefficiente di clustering alta (0,441); mentre con un numero di vicini alto (301), la media del coefficiente di clustering diminuisce (0,036).

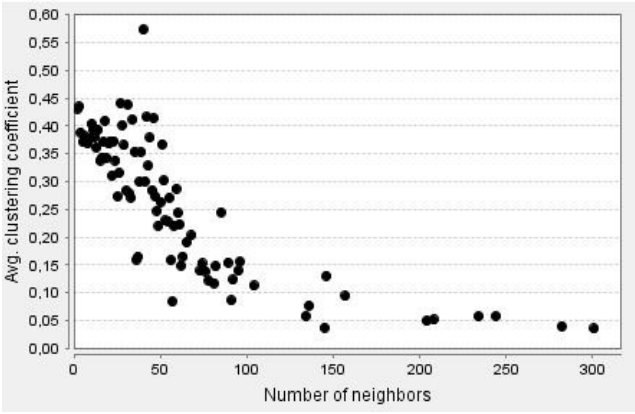

Figura 26: media del coefficiente di clustering rete foursquare

Il grafico (figura 27), in scala lineare, riguardante la Closeness Centrality mostra che la maggior parte dei nodi sono compresi nell'intervallo di valori [0.118; 0,371] e hanno un numero di vicini compreso nell'intervallo [1;96]. Questo parametro fornisce la distanza di un nodo da tutti gli altri nodi. Al contrario della Degree Centrality, questa metrica necessita di una visione globale della rete e non è quindi limitata alla visione locale dei singoli nodi.

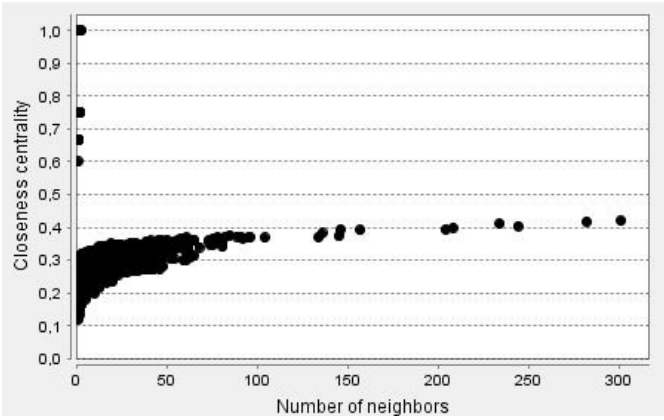

Figura 27: Closeness rete Foursquare

Il grafico (figura 28), in scala lineare, riguardante la Betweenness Centrality mostra che la maggior parte dei nodi sono compresi

nell'intervallo di valori tra [0.00;0.035] e hanno un numero di vicini compreso nell'intervallo [1;202]. La betweenness centrality si basa sull'osservazione che, se un nodo fa parte di molti cammini minimi, allora è un nodo che riveste una posizione importante nel dominio del sistema che stiamo rappresentando come rete.

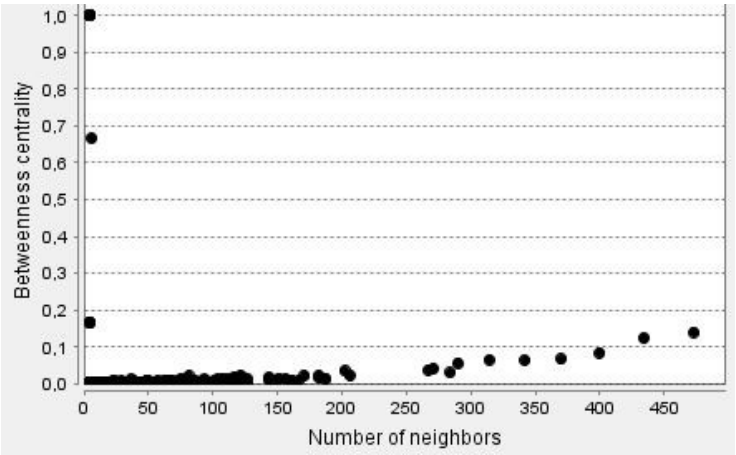

Figura 28: Betweeness rete Foursquare

Per quanto riguarda il parametro della *Node degree distribution* (figura 29), il grafico che emerge, in scala lineare, è un ramo di iperbole, in quanto si nota una relazione di proporzionalità inversa: al diminuire del numero dei nodi, aumenta il grado.

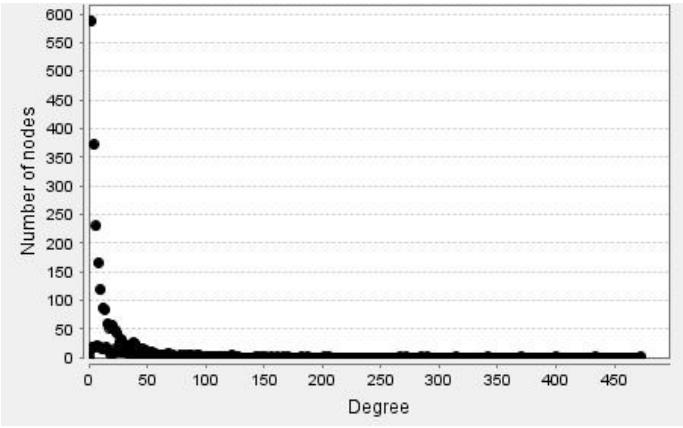

Figura 29: Node degree distribution rete Foursquare

Per quanto riguarda la *Neighboorhood connectivity* (figura 30), in scala lineare, il grafico mostra che si ha una forte densità di connettività

nell'intervallo [31,961; 54,391], con un numero di vicini basso [2;45].

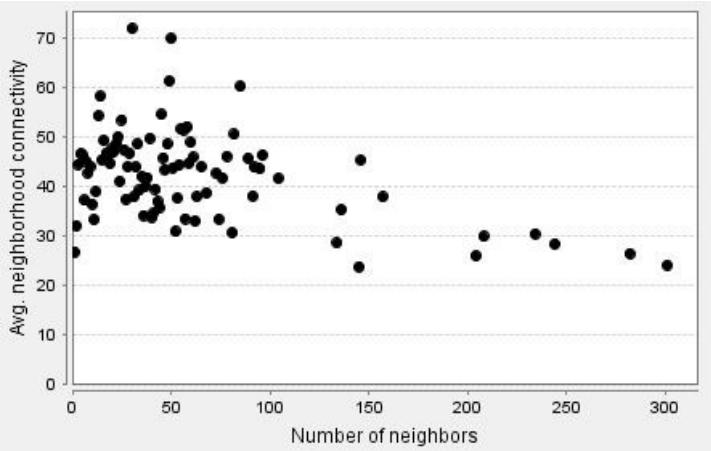

Figura 30: Neighboorhood connectivity rete Foursquare

2.2 Rete Facebook

La rete di Facebook è costituita da un campione di utenti. Gli archi identificano interazioni, dove un utente A posta sulla bacheca di un utente B, non amicizie, nell'arco di un anno. Gli archi duplicati sono stati rimossi.
Le informazioni base di tale rete sono le seguenti:

- rete non diretta;
- |V|= 8810;
- |E|= 68750;
- fonte: www.facebook.com.

Prima di addentrarci nella analisi della rete, è necessario dare qualche informazione su questo social network.
Facebook nasce nel 2004; inizialmente era stato pensato per connettere gli studenti dell'università di Harvard, poi è stato progressivamente aperto alle più importanti università americane e infine a tutti. In Italia è diventato il social network per eccellenza a partire dal settembre 2008.
Le dinamiche di utilizzo di questa piattaforma sono organizzate secondo il modello profilo/amici/condivisione. Il punto di partenza è senza dubbio la pagina del profilo che funge da identità digitale e attraverso cui si creano contatti sociali con amici, conoscenti ma anche estranei. Lo scopo dell'utilizzo di questa rete è senza dubbio la condivisione di contenuti testuali o altre forme mediali, per esprimere idee, pensieri, stati d'animo e così via.
I motivi che portano all'utilizzo di Facebook sono essenzialmente tre: il suggerimento da parte di amici, la popolarità fra i propri contatti e la conseguente curiosità e infine la voglia di mantenere contatti con altri amici.

2.2.1 Analisi

La struttura della rete è composta da una componente gigante e, vicino a questa, delle componenti connesse che sono formate da un minimo di due

nodi e un arco a un massimo di sette nodi e sette archi. Inoltre ci sono molti nodi isolati.

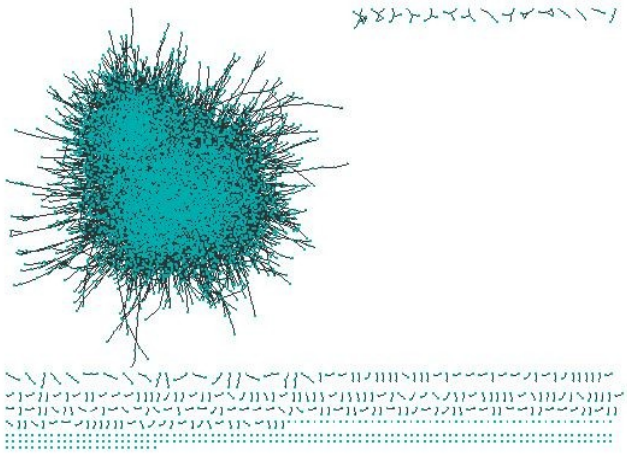

Figura 31: struttura rete Facebook

I dati elaborati dal programma sulle caratteristiche della rete sono quelli in

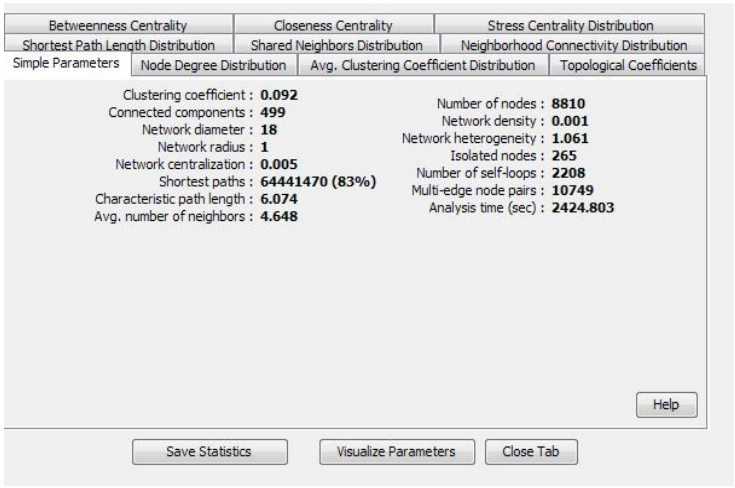

Figura 32: dati rete Facebook

Come prima misura è stato analizzato il coefficiente di clustering; il valore è di 0,092. In questo caso il valore emerso è molto basso, tendente allo 0; ciò implica che ci sono pochissime connessioni tra i nodi vicini.

La seconda misura riguarda le componenti connesse; la struttura di rete in figura 10 ci conferma il numero elevato (499).

Il diametro della rete ha un valore pari a 18.
Guardando questi parametri, si può escludere che questa rete sia una rete canonica. Infatti per considerare una rete canonica, essa deve essere poco frammentata, avere un diametro piccolo e un alto valore di coefficiente di clustering.

Il *network radius* ha un valore pari a 1.

La *Network Centralization* ha un valore pari a 0,005. Questo valore molto basso mostra che i nodi non fanno capo tutti a un unico centro; è ipotizzabile che avvenga ciò perché ogni nodo è il centro della propria rete e unendole, si perde l'unicità del centro.

La misura dello *shortest paths* è di 64441470 (83%); ciò implica che i cammini esistenti sono l'83% dei cammini possibili.

La misura riguardante la *characteristic path lenght,* ossia la lunghezza dei cammini, ha un valore pari a 6.074. Confrontando questo valore con il grafico relativo allo *Shortest path distribution,* si nota che la moda è 6. Ciò conferma la teoria dei gradi di separazione.
Il grafico presenta una distribuzione gaussiana. Per distribuzione gaussiana si intende una distribuzione di probabilità continua che spesso è usata come prima approssimazione per descrivere variabili casuali a valori reali che tendono a concentrarsi attorno a un singolo valore medio. Il grafico che emerge (figura 33) è simmetrico e ha una forma a campana, nota come campana di Gauss.

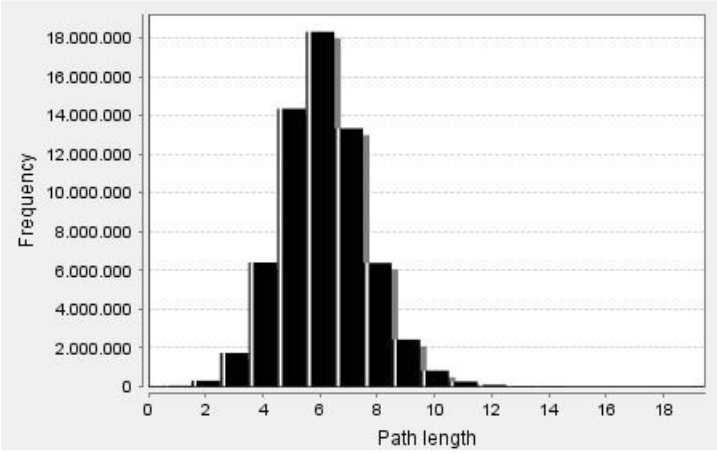

Figura 33: Shortest path distribution rete Facebook

La misura relativa all'*Avg. Numbers of neighbors* ha un valore pari al 4.648, ovvero ogni nodo ha intorno a sé circa quel numero di vicini.

La *Network density* ha un valore pari a 0,001.

Come visto dalla struttura della rete sono presenti molti nodi isolati che nella successiva analisi saranno scartati.

Infine un'altra caratteristica da evidenziare è il *number of self loops* cha ha valore di 2208; sono post sulla bacheca di un utente da parte dell'utente stesso.

Analizzando la media del coefficiente di clustering (figura 34) , con gli assi impostati su scala lineare, si nota una funzione logaritmica: all'aumentare del numero dei vicini, diminuisce la media del coefficiente di clustering.

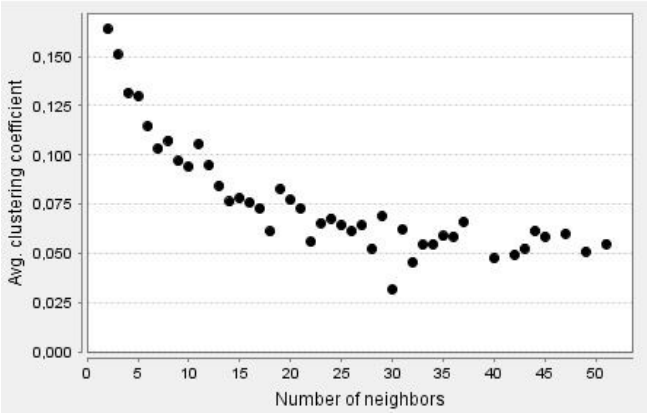

Figura 34: media coefficiente di clustering rete Facebook

Il grafico (figura 35), in scala lineare, riguardante la Closeness Centrality mostra che la maggior parte dei nodi sono compresi nell'intervallo di valori [0.082;0.222] e hanno un numero di vicini compreso nell'intervallo [1;37].

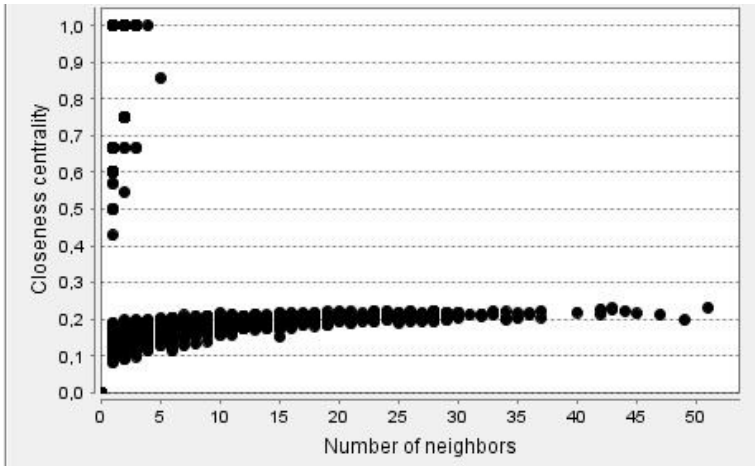

Figura 35: Closeness rete Facebook

Il grafico (figura 36), in scala lineare, della betweenness mostra che la maggior parte dei nodi sono compresi nell'intervallo [0.00;0.026] e hanno un numero di vicini compreso nell'intervallo [0;211].

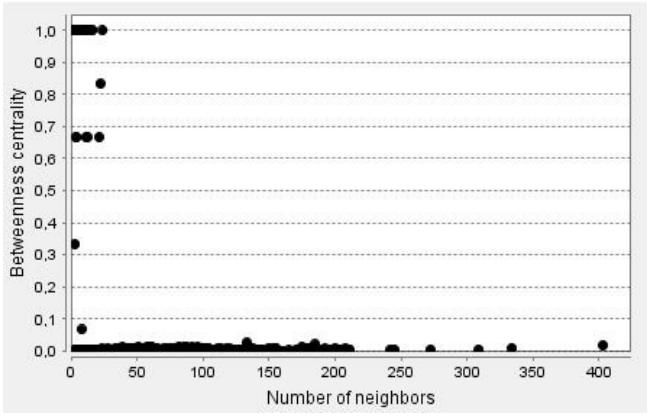

Figura 36: Betweeness rete Facebook

Per quanto riguarda la *Node Degree Distribution* (figura 37)*,* il grafico che emerge è un ramo di iperbole , in quanto si può notare una relazione di proporzionalità inversa: al diminuire del numero di nodi, aumenta il grado.

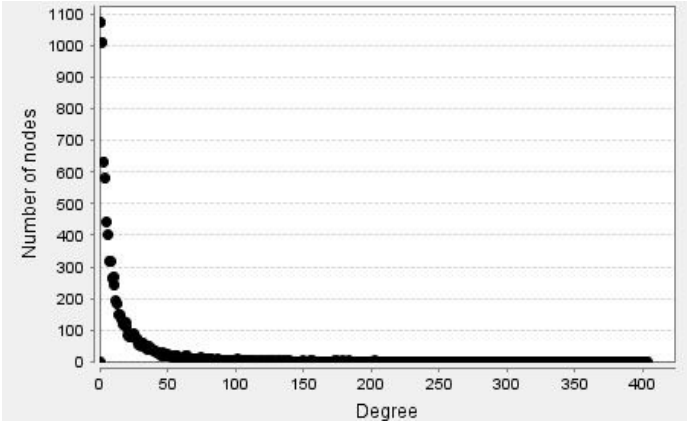

Figura 37: Node degree distribution rete Facebook

Guardando il grafico inerente la *Neighboorhod connectivity* (figura 38)*,* in scala logaritmica, si visualizza una retta , che presenta una disposizione maggiore di punti nell'intervallo [9;51] del numero dei vicini.

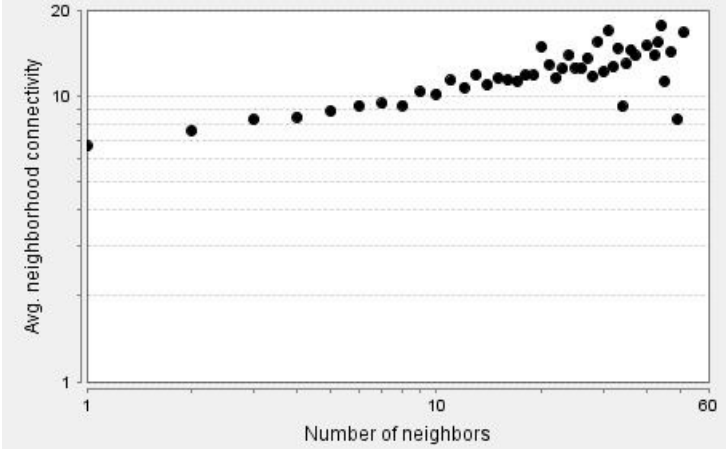

Figura 38: Neighborhood connectivity rete Facebook

2.3. Last.fm

La rete di Last.fm è costituita su un campione di utenti del servizio omonimo; i nodi rappresentano gli utenti e gli archi sono le relazioni tra gli utenti.

Last.fm è un social network musicale che ha una duplice funzione:

- per gli ascoltatori è una web radio che consiglia nuova musica sulla base degli altri brani ascoltati e graditi (sia in streaming sia sul pc tramiti appositi plugin),
- per i musicisti è uno spazio dove caricare musica che sarà fatta ascoltare agli utenti in target con il genere proposto.

Quando molti utenti del social network cominciano ad ascoltare un'artista che non è ancora presente sul database di last.fm, il social network provvede a creare una pagina dedicata in automatico.

Le informazioni di base di tale rete sono le seguenti:

- rete non diretta;
- |V|= 75969;
- |E|= 389639;
- fonte: www.lastfm.com;
- 1 componente gigante;
- 22 componenti connesse;
- Average Clustering ha un valore di 0,164: ciò implica che ci sono poche connessioni tra nodi vicini;
- Average degree ha un valore di 10,257.

In figura 39 viene riportato il grafico dell'Average degree.

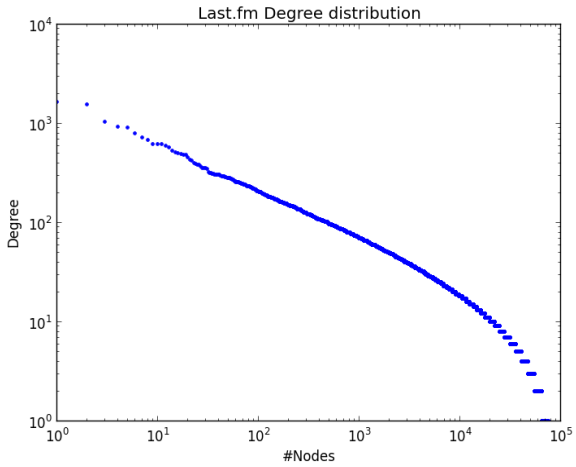

Figura 39: Degree distribution rete Last.fm

Non è stato possibile visualizzare né la struttura di rete né calcolare le altre misure di rete.

Questo è dovuto al fatto che per il software Cytoscape la rete in questione ha una dimensione troppo grande.

3. Metodologia

In questo lavoro si vuole estrarre conoscenza dalla rete al fine di poter fare previsioni su collegamenti futuri. A tale fine viene adottato il modello supervisionato, che garantisce alte performances rispetto ad un approccio non supervisionato. Il modo più naturale per utilizzare questo modello è quello di usare un classificatore su un set di attributi. Nel nostro caso sono stati usati sei classificatori appartenenti a tre famiglie (trees, SVM, Bayesiani).

È stata eseguita, infine, un'analisi comparativa su i modelli che i classificatori hanno elaborato.

3.1 Approccio supervisionato

In Data Mining si definisce "approccio supervisionato" un modello risolutivo per un particolare task in cui, una volta definiti i dati di input e output, viene lasciato al sistema il compito di apprendere la funzione che associ ad ogni dato in ingresso il suo risultato finale. Come facilmente intuibile, il buon funzionamento di questi algoritmi dipende in modo significativo da i dati in ingresso.

Diverse problemi, nell'ambito del Data Mining, possono trarre benefici da soluzioni che coinvolgano l'adozione di più modelli di analisi combinati in un approccio congiunto.

Nel nostro caso, affiancando classificazione ed algoritmi di Link Prediction non supervisionati, si punta ad ottenere un miglioramento delle performances di questi ultimi.

3.2 Algoritmi classificatori

Esistono diversi algoritmi di classificazione per l'apprendimento supervisionato. Anche se le loro prestazioni sono comparabili, alcuni di solito funzionano meglio di altri per un specifico set di dati .

In questo lavoro sono stati comparati sei algoritmi di classificazione che possiamo suddividere in tre famiglie:

- Decision tree: Bagging, C 4.5 (in weka viene chiamato J48) e Random Forest;
- SVM: è stato usato l'algoritmo SVM;

- Classificatori bayesiani: Naive Bayes Simple e BayesNet.

Gli algoritmi in questione sono stati utilizzati tramite la libreria di Weka.
Le prestazioni dei classificatori sono state comparate usando due parametri di rendimento: Accuratezza e AUROC (Area Under ROC Curve).

3.3 Cross Validation, Training e Test set

Esistono varie soluzioni per valutare le prestazioni di un modello di data mining. Per convalidare un modello in modo corretto è importante comprenderne la qualità e le caratteristiche, passo necessario, prima di renderlo disponibile in un ambiente di produzione.

Per valutare le caratteristiche e la qualità di un modello di classificazione, uno degli approcci possibili è quello di separare i dati in Training e Test Set al fine di valutarne l'accuratezza delle stime fornite.
Per Training Set si intende un insieme di dati che vengono utilizzati per addestrare un sistema supervisionato; spesso consiste in un vettore di input a cui è associata una risposta o una determinata classificazione. Una volta eseguito, l'algoritmo apprende, in base alla risposta e alla classificazione, quali caratteristiche discriminano gli elementi appartenenti alle differenti categorie. Una volta effettuata la fase di apprendimento, la correttezza dell'algoritmo viene stimata eseguendo lo stesso su un test set costruito su dati non usati in fase di apprendimento per evitare scenari di overfitting.
Si parla di overfitting quando il modello consente un'ottima classificazione sul training set, ma una pessima classificazione del test set. Come conseguenza, il modello non riesce a generalizzare poiché è basato su peculiarità specifiche del training set che non si ritrovano nel test set.
Le principali cause dell'overfitting sono:

- rumore;
- ridotta dimensione del training set.

La separazione dei dati in Training Set e Test Set rappresenta una parte importante della valutazione dei modelli di data mining. In genere, quando si separa un set di dati di Training Set e di Test Set, la maggior parte dei

dati viene usata per il training e una parte più piccola per il test.

Nei nostri test i dataset utilizzati sono stati suddivisi nel seguente modo:

- 90%Training Set;
- 10% Test Set.

Uno dei principali metodi di test e validazione di modelli di classificazione è rappresentato dalla Cross Validation

La Cross Validation è una tecnica statistica frequentemente adottata in letteratura. In particolare la k-fold-cross-validation consiste nella suddivisione del dataset in k parti di uguale dimensione: ad ogni passo una delle k parti viene usata per allenare il modello e le restanti k-1 per testarne l'accuratezza, evitando quindi problemi di overfitting ma anche di campionamento asimmetrico del dataset, tipico della suddivisione del dataset in due sole parti. In generale, quindi, si suddivide il campione osservato in gruppi di egual dimensione, si esclude iterativamente un gruppo alla volta e si cerca di classificarlo correttamente con il modello addestrato sui gruppi non esclusi.

In questa tesi abbiamo deciso di valutare i modelli costruiti impostando il valore di k pari a 10.

3.4 Metodologia di valutazione dei risultati

Per valutare la bontà del modello vengono utilizzate alcune metriche di performance. In 3.4.1 viene presentata la teoria che sta dietro alla matrice di confusione e poi vengono introdotte le curve ROC (3.4.2) e la misura di AUROC (3.4.3).

3.4.1 La matrice di confusione

Per valutare le performance di un modello predittivo si tiene traccia dei record da questi correttamente e non correttamente predetti. Tali informazioni possono essere espresse utilizzando una matrice di confusione (figura 40).

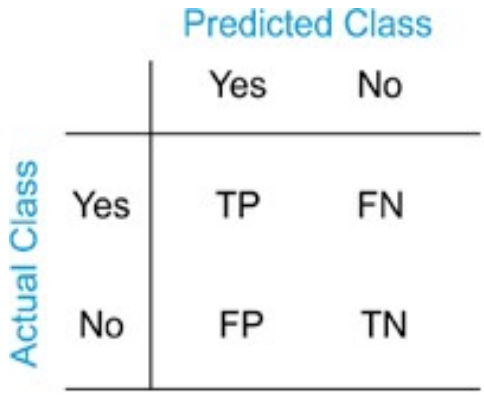

Figura 40: Matrice di confusione

Le colonne si riferiscono alla classe prevista, cioè la classe assegnata dal classificatore, mentre le righe si riferiscono alla classe effettiva.

Gli esiti predetti dal classificatore binario vengono indicati con positivi "p" e negativi "n". sono possibili quattro valori a seconda del valore di soglia:

Veri positivi (TP): record appartenenti alla classe positiva correttamente classificati come positivi.

Veri negativi (TN): record appartenenti alla classe negativa e sono stati classificati correttamente come negativi.

Falsi negativi (FN): record appartenenti alla classe positiva ma erroneamente identificati come negativi.

Falsi positivi(FP): record appartenenti alla classe negativa ma erroneamente classificati come positivi.

Sulla base di tale matrice è possibile calcolare alcune misure atte a valutare le performance di un modello.

L'accuratezza è la metrica maggiormente utilizzata per sintetizzare l'informazione da una matrice di confusione ed è data dal rapporto tra il numero di predizioni corrette e il numero totale delle predizioni. Il calcolo dell'accuratezza è:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Nel caso in cui le classi non sono bilanciate, cioè la maggior parte dei record appartengono ad una classe, l'accuratezza non è una misura adeguata.

La Precision e Recall sono due metriche utilizzate nelle applicazioni in cui la corretta classificazione dei record della classe positiva riveste una maggior importanza.

La Precision misura la frazione dei record risultati effettivamente positivi tra tutti quelli classificati come tali. I valori elevati indicano che pochi record della classe negativa sono stati erroneamente classificati come positivi.
Il calcolo della Precision è:

$$\text{Precision, } p = \frac{TP}{TP + FP}$$

La Recall misura la frazione di record positivi correttamente classificati. I valori elevati indicano che pochi record della classe positiva sono stati erroneamente classificati come negativi.
Il calcolo della Recall è :

$$\text{Recall, } r = \frac{TP}{TP + FN}$$

La Precision è uguale a 1 se tutti i record positivi sono stati effettivamente individuati.
La Recall è uguale a 1 se non ci sono falsi positivi.
Se Precision e Recall valgono entrambi 1 le classi predette coincidono con quelle reali.

Una metrica che riassume Precision e Recall è denominata F-measure. Il calcolo della F-measure è:

$$\text{F-measure}, \text{F} = \frac{2rp}{r+p} = \frac{2 \times TP}{2 \times TP + FP + FN}$$

La F-measure rappresenta la media armonica tra Precision e Recall; la media armonica tra due numeri x e y tende a essere vicina al più piccolo dei due numeri. Se la media armonica è elevata significa che sia Precision, sia Recall lo sono e quindi non si sono verificati né falsi negativi né falsi positivi.

3.4.2 Curve ROC

Le curve Receiver Operating Characteristic, o più semplicemente ROC, sono un diagramma grafico che illustra le prestazioni di un classificatore binario.

Le curve ROC furono utilizzate per la prima volta da alcuni ingegneri elettrici, che durante la seconda guerra mondiale, avevano come compito la localizzazione dei nemici mediante l'uso del radar durante le battaglie. Recentemente, invece, le curve ROC sono utilizzate in machine learning e data mining, medicina, radiologia, psicologia, veterinaria e in molti altri ambiti.

Oltre alla matrice di confusione, per disegnare una curva ROC, sono necessari due misure: TPR e FPR.

TPR (True Positive Rate) esprime la proporzione di veri positivi rispetto al numero totale di positivi effettivi. E' definita anche sensibilità.

$$\text{Recall}, \text{r} = \frac{TP}{TP + FN}$$

FPR (False Positive Rate) esprime la proporzione di falsi positivi rispetto al numero totale di negativi effettivi.

$$FPR = \frac{FP}{N} = \frac{FP}{(FP + TN)}$$

La relazione tra le suddette metriche può essere rappresentata attraverso una linea che si ottiene riportando in un sistema di assi cartesiani i valori di FPR e TPR sull'asse delle ascisse e su quello delle ordinate rispettivamente. Un test perfetto è rappresentato da una curva ROC che passa per l'angolo superiore sinistro degli assi cartesiani (massima sensibilità).

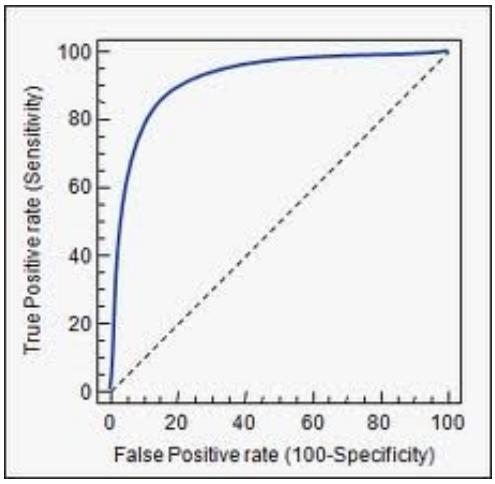

Figura 41: Curva Roc

3.4.3 Area sottesa alla curva ROC

Uno degli indici più utilizzati per valutare la bontà della regola di classificazione è l'AUC (Area under the ROC curve). Il calcolo dell'AUC per una curva empirica, cioè ottenuta da un campione finito, può essere effettuato semplicemente connettendo i diversi punti del ROC plot all'asse delle ascisse con segmenti verticali e sommando le aree dei risultanti poligoni generati nella zona sottostante.

I principali vantaggi rispetto ad altri metodi di valutazione sono i seguenti:

- è insensibile ai data set sbilanciati,
- essa valuta tutti i punti di cut-off, dando una migliore comprensione su come il classificatore sia in grado di separare le classi.

4. Risultati senza Comunità

In questo capitolo sono riportati i risultati sperimentali ottenuti a seguito dei test effettuati sui dataset precedentemente introdotti. In 4.1 si presentano i criteri di organizzazione degli esperimenti, con la suddivisione tra previsioni “non filtrate” e “filtrate”; in 4.2 è definito l'insieme delle feature usate; in 4.3 sono fornite alcune note sulle motivazioni che ci hanno spinto a non analizzare alcuni classificatori su alcune reti. Sono quindi riportati i risultati ottenuti separatamente da ciascuna famiglia dei classificatori al fine di fornirne un’analisi comparativa.

4.1 Criteri di organizzazione delle analisi

L’approccio perseguito è stato quello di creare un modello supervisionato definendo un nutrito set features.

I classificatori sono stati costruiti adottando in due differenti set di dati per ogni rete:

- Previsione non filtrata: le features sono state calcolate per tutte le potenziali coppie di nodi nella rete.
- Previsione filtrata: si riduce il numero di predizioni e si restituiscono solo i valori delle features per le coppie di nodi aventi almeno un vicino in comune.

Al fine di rendere l'analisi più agevole ci siamo concentrati sul particolare caso di analisi definito “a classi bilanciate”: tutti i modelli di classificazione analizzati sono stati costruiti a partire dallo stesso numero di istanze positive (feature estratte da coppie di nodi connesse tramite un link) e negative (feature estratte da coppie di nodi non connesse da alcuni link).

4.2 Features Set

La scelta di un set di features adeguate è la parte più critica. Per il link prediction è necessario scegliere le caratteristiche che rappresentino una qualche forma di vicinanza tra le coppie di vertici.

In questa analisi, oltre a tali caratteristiche, sono state scelte quelle che nascono dalla topologia della rete. Sono state scelte 12 caratteristiche.

La rete Facebook (filtrata e non filtrata) e la rete Foursquare-Osaka (filtrata e non filtrata) hanno lo stesso set di features per le analisi. Il set 5 è definito baseline, perché le features che ne fanno parte sono inerenti al calcolo dello score del numero dei vicini.
Features set:

- Set 0: Jaccard, Common Neighbors, Adamic Adar, Degree, Betweenness, Closeness, Eigenvector, PageRank, Hub, Authority, Cluster e Triangles;
- Set 1: Degree, Betweenness, Closeness, Eigenvector, Pagerank, Hub, Authority, Cluster e Triangles;
- Set 2: Jaccard, Common Neighbors, Adamic Adar, Degree, Betweenness, Closeness, Eigenvector, Cluster e Triangles;
- Set 3: Degree, Betweenness, Closeness, Eigenvector, Cluster e Triangles;
- Set 4: Jaccard, Common Neighbors, Adamic Adar, Pagerank, Hub, Authority;
- Set 5: Jaccard, Adamic Adar e Common Neighbors (baseline);

Nella rete Last.fm, a causa della sua dimensione, le misure di centralità sono state rimosse. Rimangono quindi le features inerenti ai parametri di link prediction (Jaccard, Adamic Adar e Common Neighbors), ai parametri inerenti al ranking (Pagerank, Hub e Autorithy), al Traingles e al Cluster. In questo caso la previsione su cui si lavora è filtrata.
Features Set:

- Set 0: Jaccard, Common Neighbors, Adamic Adar, PageRank, Hub, Authority, Triangles e Cluster;
- Set 1: PageRank, Hub, Authority, Triangles e Cluster;
- Set 2: Jaccard, Common Neighbors, Adamic Adar, PageRank, Hub, Authorithy;
- Set 3: Jaccard, Common Neighbors, Adamic Adar, Triangles e Cluster;
- Set 4: Triangles e Cluster;
- Set 5: Jaccard, Common Neighbors e Adamic Adar (baseline).

4.3 Classificatori selezionati

In alcune reti non è stato possibile analizzare i risultati di alcuni classificatori, poiché i dati in ingresso su cui avrebbero dovuto lavorare erano troppo grandi per la loro implementazione in weka. I classificatori scartati per tale motivo sono:

- BayesNet sulla rete Facebook non filtrata;
- BayesNet sulla rete Foursquare-Osaka non filtrata;
- Bagging, SVM, Naive Bayes Simple e BayesNet sulla rete Last.fm;

4.4 Interpretazione tabelle

Laddove è riportata in modo tabellare, l'analisi numerica delle performance per ogni famiglia di classificatori sono stati specificati:

- tramite il valore tra parentesi l'accuratezza;
- tramite il valore senza le parentesi l'AUROC;
- in corsivo, il valore minimo della performance del classificatore;
- in grassetto, il valore massimo della performance del classificatore.

4.5 Risultati della rete Facebook

Vengono riportati i risultati sia della rete Facebook filtrata sia della rete Facebook non filtrata per ogni famiglia di classificatore.

Decision tree su rete Facebook filtrata

Feature Set	Bagging	J48	Random Forest
Set 0	0,863 (0,779)	0,806 (0,75)	0,821 (0,742)
Set 1	0,738 (0,671)	*0,691 (0,642)*	0,711 (0,646)
Set 2	**0,864 (0,78)**	0,803 (0,752)	0,821 (0,742)
Set 3	0,74 (0,67)	0,692 (0,646)	0,702 (0,641)
Set 4	0,826 (0,75)	0,757 (0,713)	0,789 (0,725)
Set 5	0,792 (0,711)	0,756 (0,713)	0,732 (0,664)

Tabella 1: risultati dei classificatori trees

In Tabella 1 si riportano i dati della tipologia dei classificatori trees sulla rete Facebook filtrata.

Il classificatore che ha il valore più alto sia di accuratezza sia di AUROC è il Bagging nel set 2 (0,78 – 0,864).

Il classificatore che il valore più basso sia di accuratezza sia di AUROC è il J48 nel set 1 (0,642 – 0,691).

I risultati dei classificatori sono buoni; l'accuratezza si attesta nell'intervallo tra il 60% e il 75%, mentre l'AUROC si attesta nell'intervallo tra il 70% e l'85%.

Il predittore Bagging, indipendentemente dal set di features usate, risulta essere sempre il migliore.

Indipendentemente dal predittore usato il set 0 e il set 2 risultano più predittivi rispetto agli altri.
Da quanto emerso in tabella 1 le features riguardanti la centralità, il coefficiente di clustering e le misure che rappresentano la vicinanza tra le coppie dei nodi (jaccard, common neighbors e adamic adar) sono le più importanti per la predizione.

SVM su rete Facebook filtrata

Feature Set	SVM
Set 0	0,711(0,711)
Set 1	*0,621 (0,621)*
Set 2	**0,712 (0,712)**
Set 3	0,623 (0,623)
Set 4	0,693) (0,693)
Set 5	0,693 (0,693)

Tabella 2: Risultati classificatore SVM

In Tabella 2 vengono riportati i dati riguardanti il classificatore SVM sulla rete Facebook filtrata.
I valori del classificatore sono buoni e i valori di accuratezza e AUROC si stanziano nell'intervallo tra il 60% (set 1 valore più basso: 0,621) e il 70% (set 2 valore più alto: 0,714).
Il set 0 e il set 2 risultano essere i set più predittivi; quindi le features riguardanti la centralità, il coefficiente di clustering e le misure che rappresentano la vicinanza tra le coppie dei nodi sono le più rilevanti per la predizione.

Bayes su rete Facebook filtrata

Feature Set	Naive Bayes	BayesNet
Set 0	0,713 (0,663)	**0,772** (0,708)
Set 1	0,589 (*0,545)*	0,632 (0,606)
Set 2	0,712 (0,672)	**0,772 (0,718)**
Set 3	*0,577* (0,546)	0,64 (0,616)
Set 4	0,74 (0,692)	0,768 (0,708)
Set 5	0,734 (0,7)	0,749 (0,688)

Tabella 3: Risultati bayesiani

In tabella 3 vengono riportati i dati inerenti i classificatori bayesiani sulla rete Facebook filtrata.

Il classificatore che ha il valore più alto di accuratezza è il BayesNet nel set 2 (0,718); sempre il classificatore BayesNet ha il valore più alto di AUROC nel set 0 e nel set 2(0,776).

Il classificatore che il valore più basso di accuratezza è il Naives Bayes Simple nel set 1 (0,545).

Il classificatore che ha il valore più basso di AUROC è il Naive Bayes Simple nel set 3 (0,577).

A parte i due casi in cui il classificatore Naives Bayes Simple non supera la soglia del 60% dell'accuratezza nel set 1 e nel set 3, i risultati sono buoni; l'intervallo dei valori si attesta tra il 60% e il 75%. Per quanto riguarda l'AUROC, a parte l'unico caso nel set 3 in cui il classificatore Naives Bayes Simple non supera la soglia del 60%, i valori sono compresi tra il 60% e il 75%.

Il predittore BayesNet, indipendentemente dalle features usate, risulta essere sempre il migliore.

I set 0,2,4,5 risultano essere più predittivi rispetto agli altri. Dai risultati emersi, invece risulta che le features nel set 1 e 3 sono poco rilevanti per la predizione. Le features più rilevanti risultano quelle inerenti alla vicinanza tra le coppie di nodi.

Analisi comparativa rete Facebook filtrata

In tabella 4 viene riportato il top classificatore di ogni famiglia.

Feature Set	Trees (Bagging)	SVM	Bayes (BayesNet)
Set 2	**0,864 (0,78)**	0,712 (0,712)	0,772 (0,718)

Tabella 4: Tabella 4 Top classificatore di ogni famiglia

Dai risultati emersi si può notare che il Bagging sia il migliore in termini di accuratezza e di curva AUROC.

In tutti i casi la miglior performance avviene nel set 2; questo dimostra, in primo luogo, che il set 2 è più predittivo rispetto agli altri. In secondo luogo, emerge che le features più rilevanti per la predizione sono gli approcci inerenti alla link prediction e le misure di centralità.

In figure 42,43, 44 sono riportate le curve ROC dei classificatori.

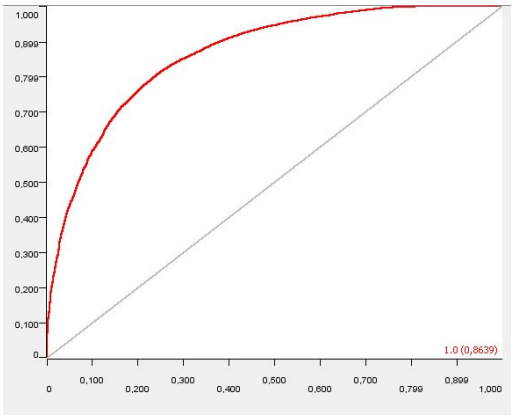

Figura 42: Curva Roc Bagging

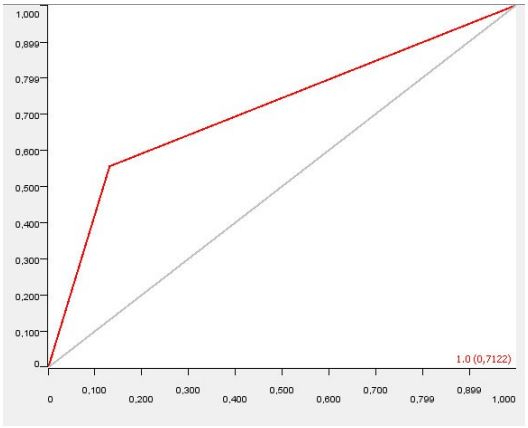

Figura 43: Curva Roc SVM

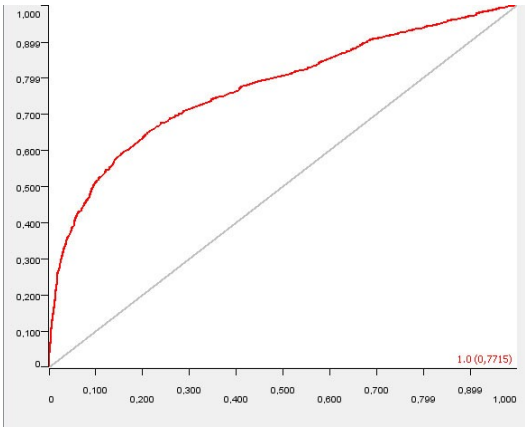

Figura 44: Curva Roc BayesNet

Trees su rete Facebook non filtrata

In tabella 5 vengono riportati i dati riguardanti i classificatori trees sulla rete Facebook non filtrata.

Feature Set	Bagging	J48	Random Forest
Set 0	0,964 (0,887)	0,939 (0,862)	0,956 (0,874)
Set 1	0,935 (0,863)	0,884 (0,816)	0,927 (0,851)
Set 2	**0,966 (0,891)**	0,937 (0,858)	0,954 (0,874)
Set 3	0,94 (0,87)	0,88 (0,813)	0,92 (0,843)
Set 4	0,947 (0,856)	0,797 (0,799)	0,933 (0,844)
Set 5	*0,773 (0,778)*	*0,773* (0,788)	0,79 (0,789)

Tabella 5: Risultati dei classificatori trees

Il classificatore che ha il valore più alto sia di accuratezza sia di AUROC è il Bagging nel set 2 (0,891 – 0,966).

Il classificatore che ha il valore più basso di accuratezza è il Bagging nel set 5 (0,778).

I classificatori che hanno i valori più bassi di AUROC sono il Bagging e il J48 nel set 5 (0,773).

I risultati dei classificatori sono ottimi; l'accuratezza si stanzia in un range tra il 75% e il 90%, mentre l'AUROC si stanzia in un range tra il 75% e il 95%.

In generale, il predittore Bagging, indipendentemente dal set di features usate, risulta essere il migliore. Dai risultati si può notare che il set 5, usando qualsiasi tipo di classificatore trees, è quello meno predittivo. Infine si può notare che le misure che rappresentano la vicinanzia tra le coppie di nodi (common neighbors, jaccard e adamic adar) risultano essere le features più rilevanti per la predizione.

SVM su rete Facebook non filtrata

Nella tabella 6 vengono riportati i dati riguardanti il classificatore SVM sulla rete Facebook non filtrata.

Feature Set	SVM
Set 0	**0,789 (0,789)**
Set 1	*0,712 (0,712)*
Set 2	**0,789 (0,789)**
Set 3	*0,712 (0,712)*
Set 4	0,788 (0,788)
Set 5	0,788 (0,788)

Tabella 6: Risultati classificatore SVM

I valori del classificatore sono buoni e i valori di accuratezza e AUROC si stanziano nell'intervallo tra il 70% (valore più basso: 0,712 nel set 1 e nel set 3) e l'80% (valore più alto: 0,789 nel set 0 e nel set 2). I set più predittivi risultano essere i set 0, 2,4,5. Ciò comporta che le features meno rilevanti risultano essere le misure di centralità e il coefficiente di clustering.

Bayes su rete Facebook non filtrata

In tabella 7 vengono riportati i dati inerenti il classificatore bayesiano sulla rete Facebook non filtrata.

Feature Set	Naive Bayes Simple
Set 0	0,864 (0,786)
Set 1	*0,767 (0,644)*
Set 2	**0,868 (0,788)**
Set 3	0,769 (0,661)
Set 4	0,853 (0,782)
Set 5	0,781 (**0,788**)

Tabella 7: Risultati del classificatore bayesiano

I valori del classificatore sono buoni. I valori di accuratezza si stanziano nell'intervallo fra il 60% (valore più basso:0,644 nel set 1) e l'80% (valore più alto: 0,788 nel set 2 e nel set 5). I valori di AUROC si stanziano nell'intervallo tra il 75% (valore più basso: 0,767 nel set 1) e il 90% (valore più alto: 0,868 nel set 2).

I set più predittivi risultano essere i set 0, 2,4. Ciò comporta che le features meno rilevanti risultano essere le misure di centralità e il coefficiente di clustering.

Analisi comparativa rete Facebook non filtrata

In tabella 8 viene riportato il top classificatore di ogni famiglia.

Feature Set	Trees (Bagging)	SVM	Bayes (Naive Bayes)
Set 2	**0,966 (0,891)**	0,789 (0,789)	0,868 (0,788)

Tabella 8: Top classificatore di ogni famiglia

Dai risultati emersi si può notare che il Bagging sia il migliore in termini di accuratezza e di curva AUROC.

In tutti i casi la miglior performance avviene nel set 2; questo dimostra, in primo luogo, che il set 2 è più predittivo rispetto agli altri. In secondo luogo, emerge che le features più rilevanti per la predizione sono le misure che rappresentano la vicinanza tra i nodi e le misure di centralità. In figura 45,46 e 47 sono riportate le curve ROC dei classificatori.

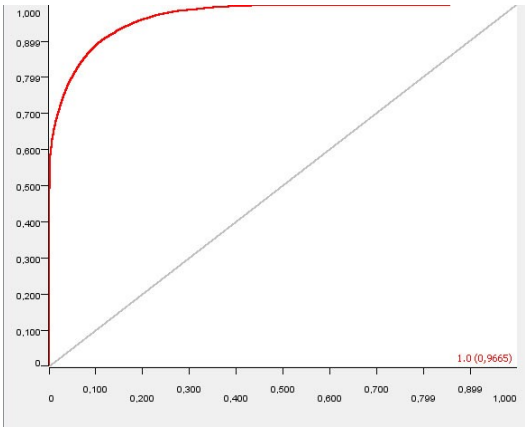

Figura 45: Curva Roc Bagging

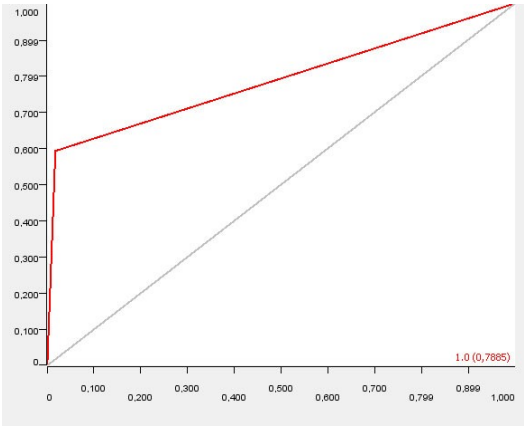

Figura 46: Curva Roc SVM

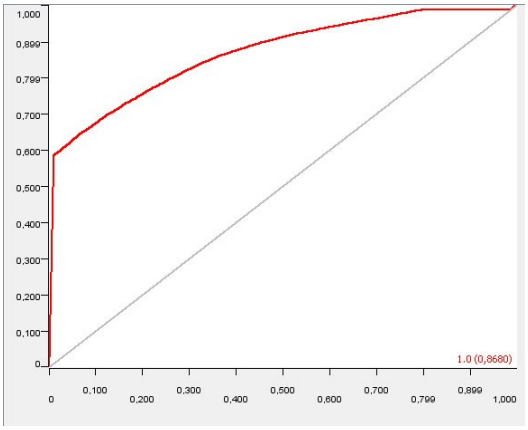

Figura 47: Curva Roc Naive Bayes

4.6 Risultati della rete Foursquare-Osaka

Vengono riportati i risultati sia della rete Foursquare-Osaka filtrata sia della rete Foursquare-Osaka non filtrata per ogni famiglia di classificatore.

Trees sulla rete Foursquare – Osaka filtrata

La tabella 9 riporta i dati della tipologia dei classificatori trees sulla rete Foursquare-Osaka filtrata.

Feature Set	Bagging	J48	Random Forest
Set 0	**0,93** (0,853)	0,87 (0,845)	0,912 (0,836)
Set 1	0,894 (0,809)	0,859 (0,781)	0,865 (0,786)
Set 2	**0,93** (**0,854**)	0,87 (0,847)	0,918 (0,844)
Set 3	0,893 (0,812)	*0,853* (0,789)	0,861 (*0,782*)
Set 4	0,922 (0,849)	0,87 (0,839)	0,901 (0,835)
Set 5	0,911 (0,835)	0,859 (0,835)	0,869 (0,8)

Tabella 9: Risultati dei classificatori trees

Il classificatore che ha il valore più alto di accuratezza è il Bagging nel set 2 (0,854).

Il classificatore che ha il valore più alto di AUROC è il Bagging nel set 0 e nel set 2 (0,93).

Il classificatore che ha il valore più basso di accuratezza è il J48 nel set 1 (0,781).

Il classificatore che il valore più basso di AUROC è il J48 nel set 3 (0,853).

I risultati dei classificatori sono ottimi; per quanto riguarda l'accuratezza i risultati sono compresi tra il 75% e il 90%, mentre per quanto riguarda l'AUROC i risultati sono compresi tra l'85% e il 95%.

In generale, il predittore Bagging risulta essere il migliore, anche cambiando il set di features.

Dai risultati emersi dai classificatori, il set 2 è quello più predittivo; in

questo caso, le features più rilevanti sono gli approcci di link prediction, le misure di centralità e il coefficiente di clustering.

SVM su rete Foursquare-Osaka filtrata

Nella tabella 10 vengono riportati i dati riguardanti il classificatore SVM sulla rete Foursquare-Osaka filtrata.

Feature Set	SVM
Set 0	0,788 (0,788)
Set 1	0,703 (0,703)
Set 2	**0,872 (0,872)**
Set 3	*0,702 (0,702)*
Set 4	0,825 (0,825)
Set 5	0,825 (0,825)

Tabella 10: Risultati classificatore SVM sulla rete Foursquare-Osaka

I valori del classificatore sono buoni e i valori di accuratezza e AUROC si stanziano nell'intervallo tra il 70% (valore più basso: 0,702 nel set 3) e il 90% (valore più alto: 0,872 nel set 2).

Dai risultati emersi dal classificatore il set 2 risulta essere più predittivo rispetto agli altri. Le features più rilevanti risultano essere le misure relative alla vicinanza tra i nodi, alle centralità e al coefficiente di clustering.

Bayes su rete Foursquare-Osaka filtrata

Nella tabella 11 vengono riportati i dati inerenti i classificatori bayesiani sulla rete Foursquare-Osaka filtrata.

Feature Set	Naives Bayes Simple	BayesNet
Set 0	0,853 (0,743)	0,892 (0,817)
Set 1	*0,71 (0,589)*	0,819 (0,731)
Set 2	0,864 (0,766)	**0,928 (0,88)**
Set 3	0,728 (*0,589*)	0,812 (0,726)
Set 4	0,87 (0,779)	0,903 (0,827)
Set 5	0,873 (0,779)	0,889 (0,816)

Tabella 11: Risultati dei classificatori bayesiani

Il classificatore che ha il valore più alto sia di accuratezza sia di AUROC è il BayesNet nel set 2 (0,88 – 0,928). Il classificatore che il valore più basso sia di accuratezza (0,589 nel set 1 e nel set 3) e di AUROC (0,71 nel set 1) è il Naives Bayes Simple. A parte i due valori bassi di accuratezza che non raggiungono il 60%, negli altri casi l'intervallo in cui rientrano i valori sono tra il 60% e l'85%. Per quanto concerne l'AUROC i valori rientrano nell'intervallo tra il 70% e il 90%.

In generale, il classificatore BayesNet risulta essere migliore rispetto al Naive Bayes Simple. Il set che risulta essere più predittivo rispetto agli altri è il 2. Le features più rilevanti sono le misure che rappresentano la vicinanza tra i nodi, le misure di centralità e il coefficiente di clustering.

Analisi comparativa rete Foursquare-Osaka filtrata

In tabella 12 viene riportato il top classificatore di ogni famiglia.

Feature Set	Trees (Bagging)	Svm	Bayes (BayesNet)
Set 2	**0,93** (0,854)	0,872 (0,872)	0,928 **(0,88)**

Tabella 12: Top classificatore di ogni famiglia

Dai risultati emersi si può notare che il Bagging sia il migliore in termini di AUROC, mentre il BayesNet risulta essere migliore in termini di accuratezza.

In tutti i casi la miglior performance avviene nel set 2; questo dimostra, in primo luogo, che il set 2 è più predittivo rispetto agli altri. In secondo luogo, emerge che le features più rilevanti per la predizione sono le misure che rappresentano la vicinanza tra i nodi e le misure di centralità.

In figura 48, 49 e 50 sono riportate le curve ROC dei classificatori.

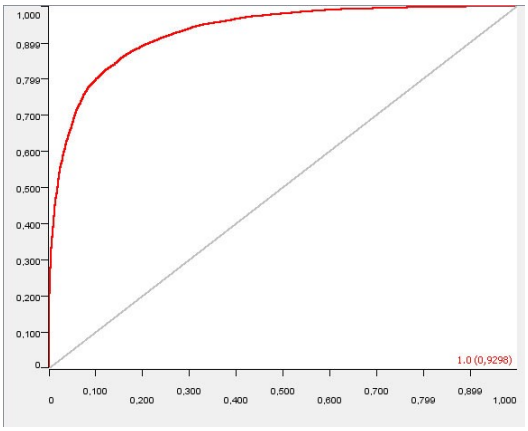

Figura 48: Curva Roc Bagging

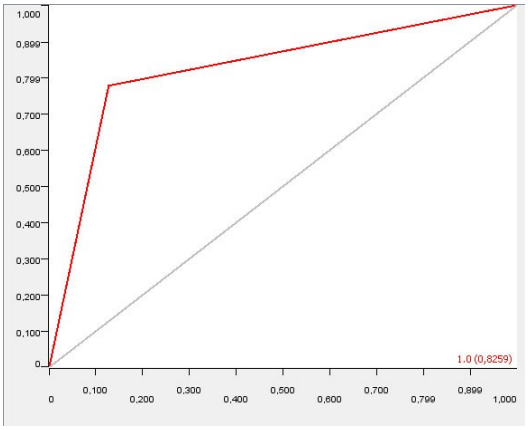

Figura 49: Curva Roc SVM

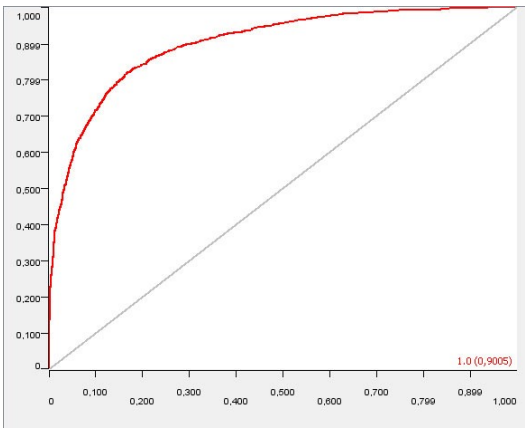

Figura 50: Curva Roc BayesNet

Trees su rete Foursquare-Osaka non filtrata

La tabella 13 riporta i dati della tipologia dei classificatori trees sulla rete Foursquare-Osaka non filtrata.

Feature Set	Bagging	J48	Random Forest
Set 0	**0,93** (0,853)	0,87 (0,845)	0,912 (0,836)
Set 1	0,894 (0,809)	0,859 (0,781)	0,865 (0,786)
Set 2	**0,93** (**0,854**)	0,87 (0,847)	0,918 (0,844)
Set 3	0,893 (0,812)	*0,853* (0,789)	0,861 (*0,782*)
Set 4	0,922 (0,849)	0,87 (0,839)	0,901 (0,835)
Set 5	0,911 (0,835)	0,859 (0,835)	0,869 (0,8)

Tabella 13: risultati dei classificatori trees

Il classificatore che ha il valore sia di accuratezza sia di AUROC più alto è il Bagging nel set 0 e nel set 2 (0,924 – 0,98).

Il classificatore che ha il valore più basso sia di accuratezza sia di AUROC è il J48 nel set 1 (0,839 – 0,886).

I risultati dei classificatori trees sono ottimi; l'accuratezza è in un' intervallo di valori tra l'80% e il 95%, mentre l'AUROC è in un intervallo di valori tra l'85% e il 98%.

In generale, il predittore Bagging risulta essere il migliore, anche cambiando il set di features.

Dai risultati emersi, il set 2 è quello più predittivo; in questo caso, le features più rilevanti sono le misure che rappresentano la vicinanza tra i nodi, le misure di centralità e il coefficiente di clustering.

SVM su rete Foursquare-Osaka non filtrata

Nella tabella 14 sono riportati i risultati del classificatore SVM sulla rete Foursquare-Osaka non filtrata.

Feature Set	SVM
Set 0	0,872 (0,872)
Set 1	0,811 (0,811)
Set 2	**0,893 (0,893)**
Set 3	*0,81 (0,81)*
Set 4	0,873 (0,873)
Set 5	0,873 (0,873)

Tabella 14: risultati del classificatore SVM

I risultati del classificatore sono ottimi; l'accuratezza e l'AUROC si stanziano tra l'80% (valore minimo: 0,81 nel set 3) e il 90% (valore massimo: 0,893 nel set 2).

Dai risultati emersi dal classificatore il set 4 e il set 5 risultano essere i più predittivi rispetto agli altri. Le features più rilevanti sono le misure che rappresentano la vicinanza tra i nodi, le misure di centralità e il coefficiente di clustering.

Bayes con rete Foursquare -Osaka non filtrata

Nella tabella 15 sono riportati i risultati del classificatore bayesiano sulla rete Foursquare-Osaka non filtrata.

Feature Set	Naive Bayes Simple
Set 0	0,933 (0,828)
Set 1	*0,83 (0,688*)
Set 2	**0,93 (0,892)**
Set 3	0,836 (0,694)
Set 4	**0,93** (**0,892**)
Set 5	0,906 (0,884)

Tabella 15: risultato classificatore bayesiano

Il classificatore Naive Bayes Simple risulta efficiente; l'accuratezza va dal

65% (valore minimo: 0,688 nel set 1) al 90% (valore massimo:0,892 nel set 2 e nel set 4). Per quanto riguarda l'AUROC, i valori vanno dal 80% (valore minimo: 0,83 nel set 1) al 95% (valore massimo: 0,93).
I set 2 e 4 risultano essere i set più predittivi.

Analisi comparativa rete Foursquare-Osaka non filtrata

In tabella 16 vengono riportati i dati del top classificatore di ogni famiglia.

Feature Set	Trees (Bagging)	SVM	Bayes (Naives Bayes)
Set 2	**0,98 (0,924)**	0,893 (0,893)	0,93 (0,892)

Tabella 16: Top-classificatore per ogni famiglia

Dai risultati emersi si può notare che il Bagging sia il migliore in termini di accuratezza e di curva AUROC.
La miglior performance dei classificatori avviene nel set 2; questo dimostra, in primo luogo, che il set 2 è più predittivo rispetto agli altri. In secondo luogo, emerge che le features più rilevanti per la predizione sono le misure che rappresentano la vicinanza tra i nodi e le misure di centralità.
In Figura 51, 52 e 53 sono riportate le curve ROC dei classificatori.

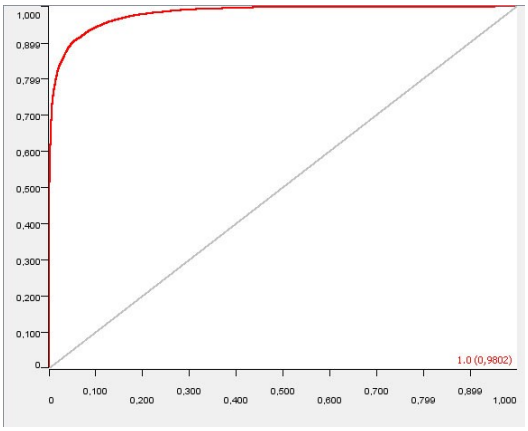

Figura 51: Curva Roc Bagging

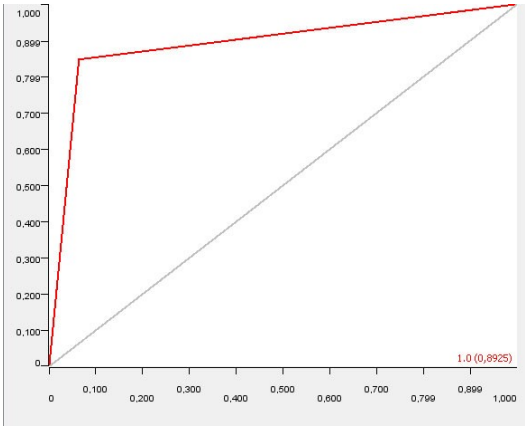

Figura 52: Curva Roc SVM

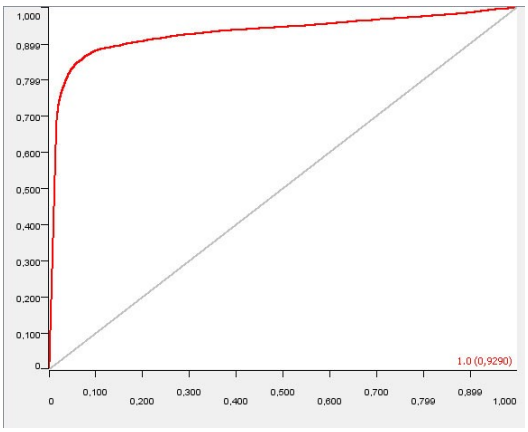

Figura 53: Curva Roc Naive Bayes

4.7 Risultati rete Last.fm

Vengono riportati in tabella 17 i risultati della rete Last.fm filtrata.

Feature Set	J48	Random Forest
Set 0	0,914 (**0,849)**	**0,924** (0,846)
Set 1	0,761 (0,699)	0,832 (0,746)
Set 2	0,896 (0,83)	0,904 (0,824)
Set 3	0,918 (0,848)	0,915 (0,834)
Set 4	0,755 (0,696)	*0,737 (0,662*)
Set 5	0,888 (0,831)	0,869 (0,795)

Tabella 17: risultati classificatori trees

Il Classificatore che ha il valore più alto di accuratezza è il J48 nel set 0 (0,849).

Il classificatore che ha il valore più alto di AUROC è il Random Forest nel set 0 (0,924).

Il classificatore che ha il valore più basso sia di accuratezza sia di AUROC è il Random Forest nel set 7 (0,662 – 0,737).

In generale, i due classificatori si equivalgono: non c'è uno che sia migliore dell'altro in termini assoluti. Nei primi tre set il classificatore Random Forest è leggermente più predittivo, mentre negli ultimi tre set è leggermente più predittivo J48. I set più predittivi sono i set 0 e 3.In questo caso le features più rilevanti sono le misure che rappresentano la vicinanza tra i nodi (jaccard, common neighbors e adamic adar) e il coefficiente di clustering.

In figura 54 e 55 sono riportate le curve Roc dei due classificatori nel set 0.

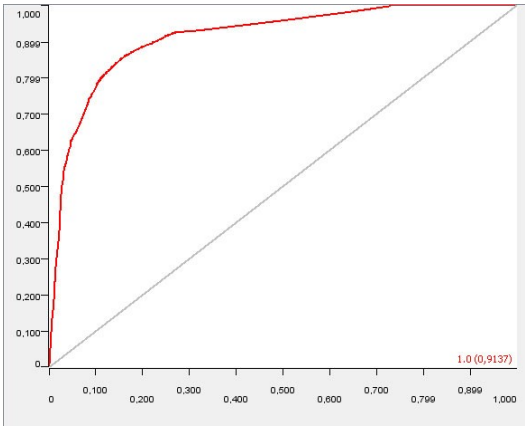

Figura 54: Curva Roc J48

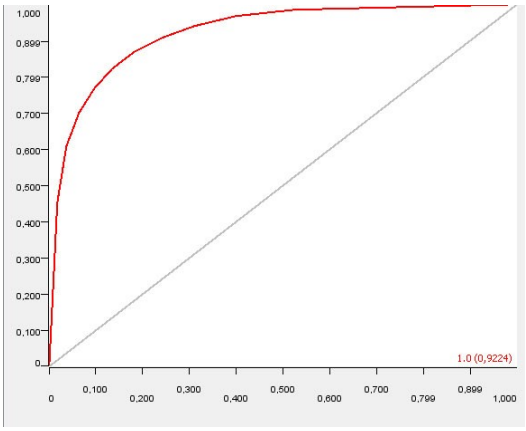

Figura 55: Curva Roc Random Forest

5. Risultati con comunità

Per migliorare i risultati ottenuti, è stata introdotta in seguito una ulteriore feature “Comunità”, che cattura il numero di comunità condivise da ciascuna coppia di nodi. Le comunità vengono estratte con l'algoritmo Demon, visto nella sezione Stato dell'arte in 3.2.1.

In 5.1 sono quindi riportati i nuovi insiemi delle feature usate; in 5.2 verranno date alcune note su i classificatori non usati in alcune reti. Infine verranno riportati i dati in tabelle suddivisi per le famiglie dei classificatori e fatta un'analisi comparativa.

5.1 Feature Set

Rispetto ai test eseguiti sui data set nel capitolo 4 è stata aggiunta la feature Comunità. Il numero delle features da 12, quindi, passa a 13.

I set di Features per le reti Facebook e Foursquare-Osaka sono gli stessi. Anche in questo caso il set 5 è quello baseline.

Features set:

- Set 0: Jaccard, Common Neighbors, Adamic Adar, Degree, Betweenness, Closeness, Eigenvector, PageRank, Hub, Authority, Triangles, Cluster, Triangles e Comunità;
- Set 1: Degree, Betweenness, Closeness, Eigenvector, PageRank, Hub, Authority, Triangles, Cluster e Comunità ;
- Set 2: Jaccard, Common Neighbors, Adamic Adar, Degree, Betweenness, Closeness, Eigenvector, PageRank, Hub, Authority e Comunità;
- Set 3: Jaccard, Common Neighbors, Adamic Adar, PageRank, Hub, Authority e Comunità;
- Set 4: Jaccard, Common Neighbors, Adamic Adar, PageRank, Hub, Authority e Comunità;
- Set 5: Jaccard, Common Neighbors, Adamic Adar e Comunità (baseline);
- Set 6: Comunità.

Per la rete Last.fm con comunità il set di features è il seguente:

- Set 0: Jaccard, Common Neighbors, Adamic Adar, PageRank, Hub, Authorithy, Cluster, Triangles e Comunità;
- Set 1: PageRank, Hub, Authority, Triangles, Cluster e Comunità;
- Set 2: Jaccard, Common Neighbors, Adamic Adar, PageRank, Hub, Authorithy e Comunità;
- Set 3: Jaccard, Adamic Adar, Common Neighbors, Triangles, Cluster e Comunità;
- Set 4: Triangles, Cluster e Comunità;
- Set 5: Jaccard, Common Neighbors, Adamic Adar e Comunità (baseline);
- Set 6: Comunità.

5.2 Classificatori selezionati

In alcune reti non è stato possibile analizzare i risultati di alcuni classificatori, poiché i dati in ingresso su cui avrebbero dovuto lavorare erano troppo grandi per la loro implementazione in weka. I classificatori non utilizzati sono:

- Naive Bayes Simple sulla rete Facebook;
- SVM sulla rete Last.fm.

5.3 Risultati della rete Facebook

Vengono riportati i risultati della rete Facebook con la features Comunità.

Trees su rete Facebook con comunità.

In tabella 18 sono riportati i risultati dei classificatori trees sulla rete Facebook con comunità.

Feature Set	Bagging	J48	Random Forest
Set 0	**0,861 (0,777**)	0,804 (0,749)	0,816 (0,735)
Set 1	0,805 (0,723)	0,76 (0,707)	0,77 (0,702)
Set 2	0,841 (0,763)	0,79 (0,746)	0,804 (0,733)
Set 3	0,849 (0,764)	0,768 (0,731)	0,81 (0,73)
Set 4	0,826 (0,747)	0,755 (0,715)	0,789 (0,719)
Set 5	0,792 (0,709)	0,753 (0,714)	0,734 (0,666)
	0,461 (0,471)	0,462 (*0,47*)	0,5 (0,5)

Tabella 18: risultati classificatori trees

Il classificatore che ha il valore più alto sia di accuratezza sia di AUROC è il Bagging (0,777 – 0,861) nel set 0.

Il classificatore che il valore più basso di accuratezza è il J48 nel set 6 (0,47).

Il classificatore che ha il valore più basso di AUROC è il Bagging nel set 6 (0,461).

A parte i valori negativi in cui si sono imbattuti i classificatori nel set 6, negli altri casi essi sono stati abbastanza efficienti; il livello di accuratezza è tra il 60% e l'80%, mentre l'AUROC è tra il 60% e l'85%.

In generale, il Bagging, indipendentemente dal set di feature usato, è il migliore.

I set 0, 2,3 risultano essere più predittivi rispetto agli altri.

Come dimostrano i risultati del set 6, la feature comunità, se usata da sola,

è poco rilevante ai fini della predizione.

SVM su rete Facebook con comunità

Nella tabella 19 sono riportati i risultati del classificatore SVM sulla rete Facebook con comunità.

Feature Set	SVM
Set 0	**0,713 (0,713)**
Set 1	0,624 (0,624)
Set 2	0,694 (0,694)
Set 3	**0,713 (0,713)**
Set 4	0,693 (0,693)
Set 5	0,693 (0,693)
Set 6	*0,5 (0,5)*

Tabella 19: risultati classificatore SVM

A parte un unico caso il cui il classificatore non raggiunge la soglia del 60%, per tutti gli altri casi risulta efficiente. L'accuratezza e l'AUROC sono tra il 60% e il 70%.

In questo caso i set più predittivo risultano essere i set 0 e 3; le features più rilevanti per la predizione sono le misure inerenti alla vicinanza tra i nodi, al pagerank e alle misure di connettività (hub, aut e coefficiente di clustering).

Il set meno predittivo risulta essere il set 6; ciò dimostra che la feature comunità, se usata da sola, non risulta essere rilevante per la predizione.

Bayes su rete Facebook con comunità

Nella tabella 20 sono riportati i risultati del classificatore bayesiano sulla rete Facebook con comunità.

Feature Set	BayesNet
Set 0	**0,77** (0,705)
Set 1	0,712 (0,622)
Set 2	0,744 (0,703)
Set 3	**0,77 (0,715**)
Set 4	0,767 (0,696)
Set 5	0,754 (0,695)
Set 6	*0,5 (0,5*)

Tabella 20: risultati del classificatore bayesiano

A parte un unico caso il cui il classificatore non raggiunge la soglia del 60%, per tutti gli altri casi risulta efficiente. L'accuratezza rientra nell'intervallo di valori tra il 60% e il 70%, mentre l'AUROC rientra in un intervallo di valori tra il 60% e l'80%.

In questo caso non c'è un vero e proprio set che incida più degli altri; si può notare che il set 6, con la feature comunità, è il meno predittivo rispetto agli altri.

Analisi comparativa Rete Facebook con comunità

In tabella 21 viene riportato il top classificatore di ogni famiglia.

Feature set	Trees (Bagging)	SVM	Bayes (BayesNet)
Set 0	**0,861 (0,777)**	0,713 (0,713)	0,77 (0,705)

Tabella 21: Top-classificatore per ogni famiglia

Dai risultati riportati si può notare che il Bagging risulta essere il miglior predittore sia in termini di accuratezza si in termine di AUROC.

La miglior performance dei classificatori avviene nel set 0, dove ci sono

tutte le features.

Confrontando i dati con il miglior top classificatore della rete Facebook filtrata in 4.5.5 si nota che i risultati sono simili e che quindi la features comunità non influisce in maniera determinante.

In figura 56,57 e 58 sono riportate le curve ROC dei classificatori.

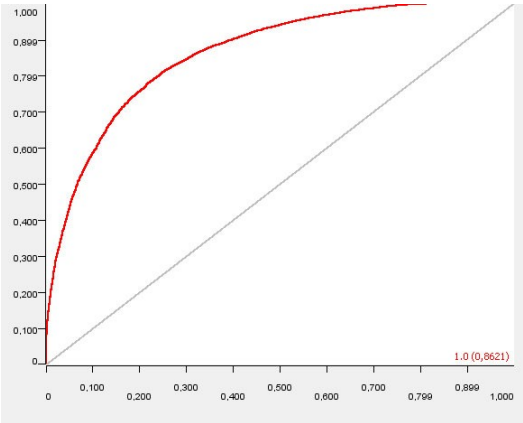

Figura 56: Curva Roc Bagging

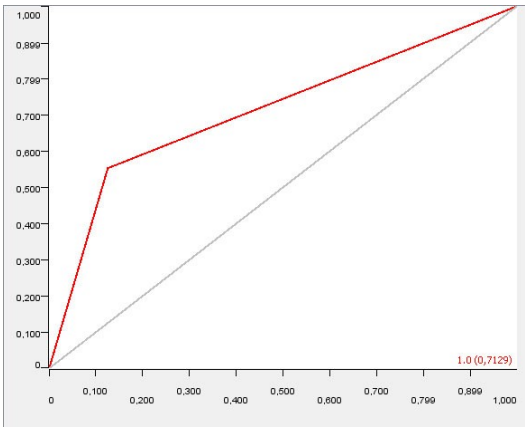

Figura 57: Curva Roc SVM

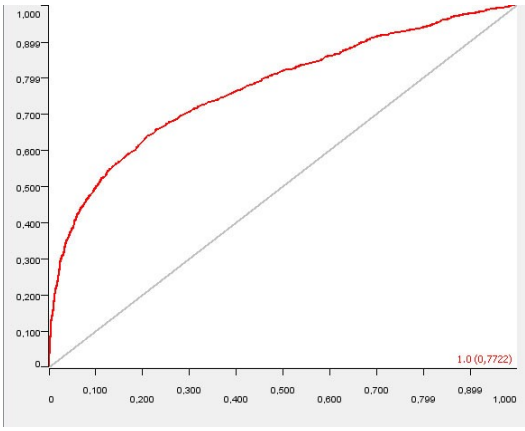

Figura 58: Curva Roc BayesNet

5.4 Risultati della rete FourSquare-Osaka con comunità

Vengono riportati i risultati della rete Foursquare-Osaka con la feature comunità.

Trees su rete Foursquare-Osaka con comunità

La tabella 22 riporta i dati della tipologia dei classificatori trees sulla rete Foursquare-Osaka con comunità.

Feature Set	Bagging	J48	Random Forest
Set 0	**0,861 (0,777**)	0,804 (0,749)	0,816 (0,735)
Set 1	0,805 (0,723)	0,76 (0,707)	0,77 (0,702)
Set 2	0,841 (0,763)	0,79 (0,746)	0,804 (0,733)
Set 3	0,849 (0,764)	0,768 (0,731)	0,81 (0,73)
Set 4	0,826 (0,747)	0,755 (0,715)	0,789 (0,719)
Set 5	0,792 (0,709)	0,753 (0,714)	0,734 (0,666)
Set 6	*0,461* (0,471)	0,462 (*0,47*)	0,5 (0,5)

Tabella 22: risultati classificatori trees

Il classificatore che ha il valore più alto sia di accuratezza sia di AUROC è il Bagging nel set 0 (0,993- 0,998).

Il classificatore che ha il valore più basso di accuratezza è il Random Forest nel set 5 (0,8).

Il classificatore che ha il valore più basso di AUROC è il J48 nel set 5 (0,857).

I classificatori trees sono molto efficienti: a livello di accuratezza, l'intervallo dei valori è tra l'80% e il 100%. A livello di AUROC, l'intervallo dei valori è tra l'85% e il 100%.

Indipendentemente dal set di features usate, il predittore Bagging risulta essere il migliore.

In linea generale, tutti i set sono molto predittivi; i valori più alti si trovano

nei set 0 e set 3.

Se comparato con il set 6 della rete Facebook in 5.3.2, il set 6 della rete Foursquare-Osaka risulta essere molto predittivo.

Le features più rilevanti risultano essere quelle inerenti alle misure di centralità.

SVM sulla rete Foursquare-Osaka con comunità

La tabella 23 riporta i dati del classificatori SVM sulla rete Foursquare-Osaka con comunità.

Feature Set	SVM
Set 0	**0,871 (0,871)**
Set 1	*0,708 (0,708)*
Set 2	0,825 (0,825)
Set 3	0,786 (0,786)
Set 4	0,826 (0,826)
Set 5	0,825 (0,825)
Set 6	**0,871 (0,871)**

Tabella 23: risultati classificatore SVM

I risultati del classificatore SVM sono ottimi; il range dei valori per l'accuratezza e l'AUROC va dal 70% (valore minimo: 0,798 nel set 1) al 90% (valore massimo:0,871 nel set 0 e nel set 6).

Il set 0 e il set 6 risultano essere i set più predittivi; la feature comunità, da sola risulta essere rilevante ai fini della predizione.

Bayes sulla rete Foursquare-Osaka con comunità

La tabella 24 riporta i dati della tipologia dei classificatori bayesiani sulla rete Foursquare-Osaka con comunità.

Features Set	Naive Bayes Simple	Bayes Net
Set 0	0,878 (0,768)	**0,921 (0,838)**
Set 1	*0,782 (0,656)*	0,848 (0,764)
Set 2	0,859 (0,752)	0,89 (0,815)
Set 3	0,862 (0,768)	0,901 (0,822)
Set 4	0,871 (0,779)	0,902 (0,827)
Set 5	0,784 (0,779)	0,893 (0,82)
Set 6	0,92 (0,787)	0,896 (0,815)

Tabella 24: risultati dei classificatori bayesiani

Il classificatore che ha il valore più alto sia di accuratezza sia di AUROC è il BayesNet nel set 0 (0,838 – 0,921).

Il classificatore che ha il valore più alto sia di accuratezza sia di AUROC è il Naive Bayes Simple nel set 1(0,656 – 0,782).

I classificatori operano efficientemente; per quanto riguarda l'accuratezza il range dei valori va dal 65% al 90%, mentre per l'AUROC il range dei valori va dal dal 75% al 95%.

Indipendentemente dal set di features usato, il classificatore BayesNet risulta essere migliore rispetto al Naive Bayes Simple.

I set più predittivi indipendentemente dal predittore usato risultano essere il set 0 e il set 6.

La feature comunità risulta essere rilevante per una buona predizione.

Analisi comparativa della rete Foursquare- Osaka con comunità

In tabella 25 viene riportato il top classificatore per ogni famiglia.

Feature set	Trees (Bagging)	SVM	Bayes (BayesNet)
Set 0	**0,861 (0,777)**	0,713 (0,713)	0,77 (0,705)

Tabella 25: Top-classificatore di ogni famiglia

Dai risultati riportati si può notare che il Bagging risulta essere il miglior predittore sia in termini di accuratezza si in termine di AUROC.

La miglior performance dei classificatori avviene nel set 0, dove ci sono tutte le features.

In Figura 59,60 e 61 sono riportate le curve ROC dei classificatori.

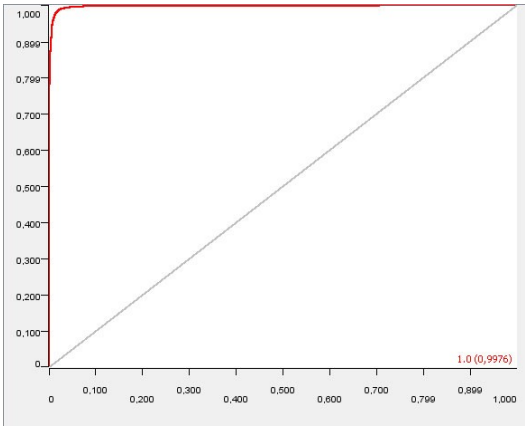

Figura 59: Curva Roc Bagging

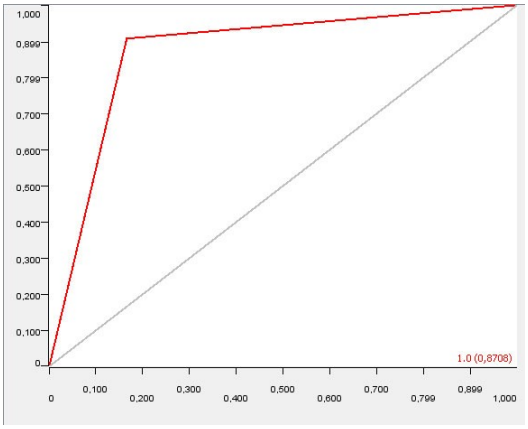

Figura 60: Curva Roc SVM

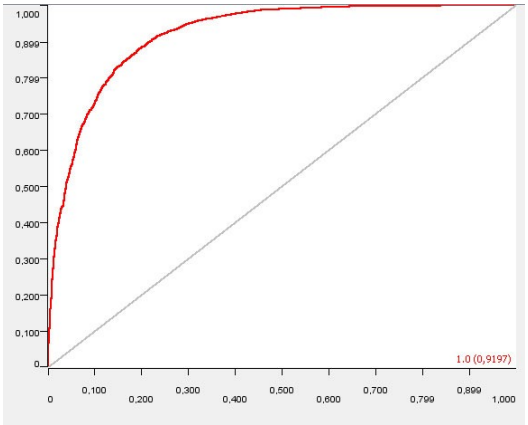

Figura 61: Curva Roc BayesNet

5.5. Risultati della rete Last.fm con comunità

Di seguito si riportano i risultati della rete Last.fm con la feature comunità.

Trees sulla rete Last.fm con comunità

La tabella 26 riporta i dati dei classificatori trees sulla rete Last.fm con comunità.

Features Set	Bagging	J48	Random Forest
Set 0	*0,462* (*0,471*)	*0,462* (*0,471*)	0,5 (0,5)
Set 1	-	0,759 (0,698)	-
Set 2	-	-	-
Set 3	-	**0,918** (**0,848**)	0,914 (0,834)
Set 4	0,777 (0,693)	0,756 (0,678)	0,778 (0,693)
Set 5	-	0,888 (0,83)	-
Set 6	0,795 (0,807)	0,795 (0,807)	0,864 (0,808)

Tabella 26: risultati classificatori trees

Con alcuni set i classificatori non hanno potuto elaborare i dati. Si può notare che il classificatore J48 risulta essere più affidabile, in quanto riesce a lavorare su quasi tutti i set. Il classificatore J48 ha il valore sia di accuratezza sia di AUROC più alto. Nei set di features in cui tutti e tre i predittori hanno potuto lavorare si può notare che Rando Forest risulta essere il migliore. I set che sono più predittivi sono il 3,4 e 6. Le features meno rilevanti per la predizione sono quelli inerenti al pagerank, all'authority e all'hub.

Bayes sulla rete Last.fm con comunità

La tabella 27 riporta i dati dei classificatori bayes sulla rete Last.fm con comunità.

Feature Set	Naive Bayes Simple	Bayes Net
Set 0	*0,461* (*0,47*)	0,5 (0,5)
Set 1	-	-
Set 2	0,868 (0,742)	-
Set 3	0,857 (0,762)	**0,91 (0,804)**
Set 4	0,65 (0,573)	0,748 (0,666)
Set 5	-	-
Set 6	0,858 (0,669)	0,864 (0,798)

Tabella 27: risultati classificatori bayesiani

Con alcuni set i classificatori non hanno potuto elaborare i dati. Si può notare che il classificatore Naive Bayes Simple risulta essere più affidabile, in quanto riesce a lavorare su quasi tutti i set. Il classificatore Naive Bayes ha il valore di AUROC più alto, mentre BayesNet ha il valore più alto di accuratezza. Nei set di features in cui tutti e tre i predittori hanno potuto lavorare si può notare che BayesNet risulta essere il migliore. I set che sono più predittivi sono il 3 e il 6. Le features più rilevanti sono quelle inerenti alle misure di vicinanza tra i nodi, al coefficiente di clustering e alla comunità.

Analisi comparativa rete Last.fm con comunità

In tabella 28 viene riportato il top-classificatore di ogni famiglia.

Feature Set	Trees (J48)	Bayes (BayesNet)
Set 3	**0,918 (0,848)**	0,91 (0,804)

Tabella 28: Top-classificatore di ogni famiglia

Dai risultati riportati si può notare che il J48 risulta essere il miglior predittore sia in termini di accuratezza si in termine di AUROC.

La miglior performance dei top-classificatori avviene nel set 3; come

conseguenza le features più rilevanti al fine della predizione sono quelle inerenti alle misure di vicinanza tra i nodi, al coefficiente di clustering e al triangles.

In figura 62 e 63 sono riportate le curve ROC dei classificatori.

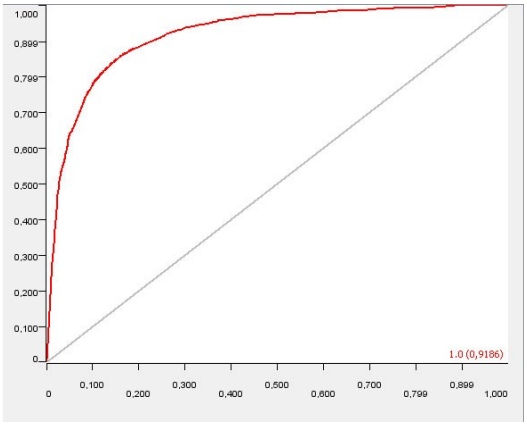

Figura 62: Curva Roc J48

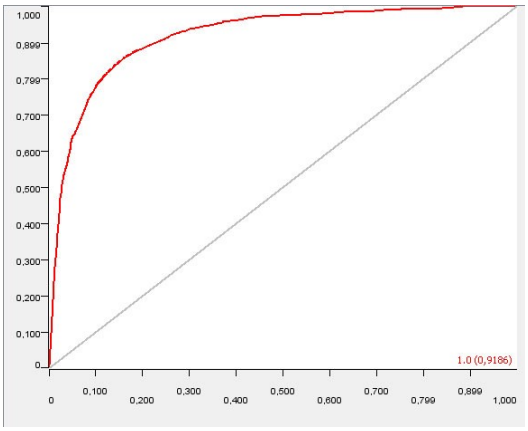

Figura 63: Curva Roc BayesNet

6.0 Discussione dei risultati

Il link prediction nei social network è un problema importante ed è molto utile per analizzare e comprendere i gruppi sociali. Tale comprensione può portare all'efficace attuazione di strumenti per identificare gruppi nascosti o per trovare membri mancanti di gruppi.

Attraverso questo lavoro, è stato dimostrato che il link prediction nelle reti sociali può avere un'alta precisione, considerando alcune caratteristiche.

È stato dimostrato che gli algoritmi di classificazione risultano essere efficaci nel predire nuovi archi futuri.

Dai risultati emerge che i classificatori della famiglia trees hanno performances migliori rispetto a quelle delle famiglie Svm e bayesiani.

Confrontando le performances tra le reti filtrate e le reti non filtrate si può notare che quelle non filtrate hanno valori migliori. Quindi, a livello di previsioni, il filtro non si dimostra molto rilevante; serve, invece, per ridurre la grandezza del numero di previsioni.

Inserendo la feature comunità, solo nel caso della rete Foursquare-Osaka si ha un netto miglioramento; nei casi delle reti Facebook e Last.fm, i valori si stanziano all'incirca su quelli delle reti filtrate.

Per quanto riguarda le classi di features che risultano essere rilevanti ai fini della predizione, si nota che:

- nel caso delle reti filtrate e non filtrate di Facebook e Foursquare-Osaka è sempre il set 2 a essere il più predittivo. Quindi, le features più rilevanti risultano essere quelle riguardanti le misure di vicinanza tra i nodi e di centralità.
- Nel caso delle reti Facebook e Foursquare-Osaka con comunità e Last.fm, è il set 0 quello più predittivo. In questo set sono presenti tutte le features.
- Infine, nel caso della rete Last.fm con comunità, il set più predittivo risulta essere il 3 e le features più rilevanti sono quelle inerenti alle misure di vicinanza, al coefficiente di clustering e al triangles.

Infine si può affermare che le misure inerenti al pagerank, all'hub e

all'authority si dimostrano poco rilevanti ai fini della predizione; mentre, importantissimi sono quelle che misurano la vicinanza tra i nodi (jaccard, adamic adar e common neighbors).

PARTE V
CONCLUSIONI

CONCLUSIONI

Dopo aver affrontato la definizione del problema nella *Parte II*, in questo ultimo capitolo si riassumono i risultati dell'analisi sperimentale affrontata nella *Parte IV.*
Dopo aver riassunto quanto è stato presentato nel corso del lavoro viene data un'analisi dei risultati. Successivamente, viene data un'analisi di possibili lavori futuri da intraprendere per sviluppare l'argomento affrontato alla luce di quanto emerso dai dati sperimentali.

1. Valutazione dei risultati ottenuti

In questo lavoro di tesi è stato introdotto il task di Link Prediction, ossia il problema di prevedere le nuove connessioni che si instaureranno fra i nodi di una rete.
Abbiamo osservato come lo studio del problema tramite l'adozione di modelli supervisionati permetta di garantire buone perfromance nell'accuratezza delle previsioni. Per garantire tali risultati sono stati affiancati alla classificazione, un processo del data mining, algoritmi non supervisionati di Link Prediction. La maniera più naturale per utilizzare questa metodologia è stata quella di creare classificatori che prendessero come input insiemi di feature rappresentanti le caratteristiche topologiche della rete. A tale fine, sei algoritmi di classificazione sono stati utilizzati appartenenti a tre famiglie (Decision Trees, SVM, Bayesiani).

Attraverso questo lavoro è stato mostrato che il problema di Link Prediction in contesti di rete sociale può essere risolto garantendo un'accuratezza molto elevata considerando un numero limitato di caratteristiche topologiche. Inoltre, si è osservato che il modello adottato, facente uso di algoritmi di classificazione, può risolvere il problema fornendo predizioni aventi alta precisione.

E' stato inoltre studiato come avere informazioni relative all'appartenenza

ad una stessa comunità dei nodi tra cui si vuol prevedere una nuova interazione influenzi l'accuratezza del modello proposto. Contrariamente a quanto atteso, l'adozione della feature "comunità" non ha portato, in generale, ad un aumento sensibile di un'accuratezza già molto buona: infatti, solo nel caso della rete Foursquare si è registrato un netto miglioramento nelle performance rasentando, nel caso dei classificatori della famiglia Decision Tree, l'accuratezza massima raggiungibile.

Guardando i risultati dei classificatori si può notare che quelli appartenenti alla tale famiglia (in particolare, il classificatore Bagging) hanno perfomance migliori rispetto alle altre famiglie.
Per quanto riguarda la rilevanza delle feature, di grande impatto sui risultati della predizione si sono dimostrate le misure di centralità e quelle inerenti la vicinanza tra i nodi (Jaccard, Adamic Adar e Common Neighbors). Al contrario, poco rilevanti ai fini della predizione si sono rivelate le misure inerenti al node ranking (PageRank, Hub e Authority).

Riassumendo, possiamo dire i maggiori contributi portati da questo lavoro di tesi sono:

Analisi di tre reti sociali;

Costruzione di modelli supervisionati, basato sulla tecnica di classificazione, per una previsione accurata di nuovi link;

Estensiva sperimentazione atta a dimostrare la bontà dell'approccio proposto.

2. Lavori futuri

Questo lavoro di tesi ha considerato il problema di Link Prediction su reti sociali.
Una interessante estensione potrebbe portare all'analisi dei modelli predittivi proposti in questo lavoro in contesti semanticamente diversi, ovvero, studiare se la previsione dei collegamenti varia in maniera significativa su reti appartenenti a domini diversi (reti biologiche, tecnologiche, World Wide Web).

Un altro aspetto da tenere in considerazione è quello che nella nostra lista di attributi non è stato considerato il dominio del tempo. Sarebbe stimolante applicare un modello supervisionato a reti osservate in diversi istanti temporali, reti per cui si ha la conoscenza del momento in cui ogni nodo e ogni arco appare nella rete. Un ulteriore aspetto potrebbe essere quello di cambiare il set di feature analizzate al fine di osservare se, con altri insiemi di caratteristiche, sia possibile ottenere predizioni aventi un'elevata accuratezza.
Infine, sarebbe da considerare questo modello per una analisi su reti multidimensionali, reti in cui possono essere presenti molteplici archi, potenzialmente aventi diversa semantica, tra ogni coppia di nodi.

BIBLIOGRAFIA

[1] M.E.J. Newman. The structure and fuction of complex networks. Siae Review, 2(45:167,2003.

[2] Albert-Lazlò Barabasi. Link, la scienza delle reti. Einaudi, 2011.

[3] Deepayan Chackrabarti and Christos Faloutsos. Graph mining: Laws, generators, and algorithms. AMC Computing Surveys, 38 March 2006.

[4] Albert- Lazlò Barabasi. Network Science. PDF version, November 2012.

[5] Ayman Farahat, Thomas Loforo, Joel C.Miller, Gregory Rae, and Lesly A.Ward. Autorithy rankings from Hits, PageRank and Salsa: existence, uniqueness and effect of inizialization. ACM SIGIR, 2001.

[6] Michele Coscia, Fosca Giannotti, and Dino Pedreschi. A classification for Community Discovery Methods in Complex Networks. Wiley Periodicals, August 2011.

[7] David Liben-Nowell and Jon Kleimberg. The link-prediction problem for social networks. Journal of the American Society for Information Science and Technology: 1019-1031, May 2007.

[8] Davide Zamperin. Strumenti open source per Data Mining. Confronto tra Weka e R. 2007.

[9] Pang-Ning Tan, Michael Steinback, Vipin Kubar. Introduction to Data Mining. Addison Wesley,ISBN 0-321-32136-7, 2006.

[10] Travers and Milgram. An experimental study of the small world problem. Sociometry, 1969.

[11] Peter Sheridan Dodds, Roby Muhammad, Duncan J.Watts. An Experimental Study of Searchin Global Social Networks. Science, Vol.301, August 2003.

[12] Paolo Rossi. Metodologie fisiche per le scienze umane. 2009.

[13] L.Page, S.Brin, R. Motwani, and T.Winograd. The pagerank citation ranking: Brinding order to the web. 1999.

[14] en.wikipedia.org/wiki/PageRank.

[15] Lise Geotoor and Cristopher P.Diehl. Link Mining: A survey. SIGKDD Explorations, 7(2).

[16] Alexandrin Popescu and Lyle H.Ungar. Statistical relational learning

for Link Prediction. IJCAI, 2003.
[17] Yunpeng Zhao, Elizaveta Levina, and Ji Zhu. Link prediction for partially observed networks. 29 January 2013.
[18] Giulio Rossetti, Michele Berlingerio, Fosca Giannotti. Scalable Link Prediction on Multidimensional networks.
[19] Zan Huang, Xin Li, and Hsinchum Chen. Link Prediction to approach to collaboritve filtering. JCDL, 2005.
[20] M.E.J. Newman. Clustering and preferential attachment in growing networks. April 2001.
[21] David M. Pennock, Gary W.Flake, Steve Lawrence, Eric J.Glover, and C.Lee Giles. Winners don't take all: Characterizing the competition for links on the web. 2002.
[22] J.H.Jones and M.S. Handcock. An assesment of preferential attachment as a mechanism for human sexual network formation. The Royal Society, 2003.
[23] Saket Navlaka and Carl Kingsford. Network archaelogy: Uncovering ancient networks from present-day interactions. September 2010.
[24] Kai Yu and Wei Chu. Stochastic relational models for discriminative Link Prediction.
[25] Chao Wang, Venu Satuluri, and Srinivisan Parthasarathy. Local probabilistic model for Link Prediction.
[26] Jeremy Kubica, Andrew Moore, David Cohn, and Jeff Schneider. A fast graph-based method for link analysis and querues.
[27] Joshua O'Madadhain, Jon Hutchins, and Pedhraic Smyth. Prediction and ranking algorithms for event-based network data. SIGKDD Explorations.
[28] Michele Berlingerio, Francesco Bonchi, Bjorn Bringman, and Aristides Gionis. Mining graph evolution rules.
[29] Cane Wing Ki Leung, Ee-Peng Lim, David Lo, and Jianshu Weng. Mining interesting link formation rules in social networks. CIKM'10, October 2010.
[30] Mohammad Al Hasan, Vineet Chaoji, Saeed Salem, and Mohammed Zaki. Link Prediction using supervised learning.
[31] Mustafa Bilgic, Galileo Mark Namata, and Lise Getoor. Combining

collective classification and Link Prediction.

[32] A. Potgieter, Kurt April, R.J.E Cooke, and I.O Osunmakinde. Temporality in Link Prediction: Understanding social complexity. Sprouts: Working Papers on Information Systems, 2007.

[33] L.Lu and T.Zhou. Link Prediction in complex networks: A survey. Physica: A: Statistical Mechanics and its applications, 390(6):1150-1170, 2011.

[34] C.A. Bliss, M.R Frank, C.M. Danforth, and P.S. Dodds. An evolutionary algorithm approach to link prediction in dynamic social networks. ArXiv preprint arXiv: 1304.6257, 2013.

[35] Z.Bao, Y.Zeng, and Y.Tay. Sonlp: Social network link prediction by principal component regression. In Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, pages 364-371. ACM, 2013.

[36] M.Pujari and R. Kanawati. Supervised rank aggregation approach for link prediction in complex network. In Proceedings of the 21st international conference companion on World Wide Web, pages 1189-1196. ACM, 2012.

[37] N.Shibata, Y. Kajikawa, and I.Sakata. Link prediction in citation networks. Journal of the American Society for Information Science and Technology, 63(1):78-85, 2012.

[38] R.N. Lichtenwalter, J.T. Lussier, and N.V. Chawla. New perspectives and methods in link prediction. In Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining, pages 243-252. ACM, 2010.

[39] S. Sounddarajan and J. Hopcroft. Using community information to improve the precision of link prediction methods. In Proceedings of the 21st international conference companion on World Wide Web, pages 607-608. ACM, 2012.

[40] X. Feng, J.Zhao, and K.Xu. Link prediction in complex networks: a clustering perpespective. The European Physical Journal B, 85(1): 1-9, 2012.

[41] K. Jahanbakhsh, V.King, and G.C. Shoja. Predicting human contacts in mobile social networks using supervised learning. In Proceedings of the Fourth Annual Workshop on Simplyfing Complex Networks for

Practitioners, pages 37-42. ACM,2012.

[42] M.Fire, R.Puzis, and Y.Elovici. Link prediction in highly fractional data sets. In Handbook of Computational Approaches to Counterterrorism, pages 283-300. Springer, 2013.

[43] Y.Xu and D.Rockmore. Feature selection for link prediction. In Proceedings of the 5th Ph. D. workshop on Information and knowledge, pages 25-32. ACM,2012.

[44] R.Lichtnwalter and N.V. Chawla. Link prediction: fair and effective evaluation. In Advances in Social Netwrok Analysis and Mining (ASONAM), 2012 IEEE/ACM International Conference on, pages 376-383. IEEE, 2012.

[45] S. Spiegel, J.Clausen, S. Albayrak, and J. Kunegis. Link prediction on evolving data using tensor factorization. In New Frontiers in Applied Data Mining, pages 100-110. Springer, 2012.

[46] P.Sarkar, D.Chakabarti, and M.Jordan. Nonparametric link prediction in dynamic networks. ArXiv preprint arXiv: 1206.6394, 2012.

[47] P.R da Silva Soares and R.Bastos Cavalcante Prudencio. Time series based link prediction. In Neural Networks (IJCNN), The 2012 International Joint Conference on, pages 1-7. IEEE, 2012.

[48] http://it.wikipedia.org/wiki/Albero_di_decisione.

[49] Nicola Fulvio Calabria. Classificazione di segnali ed immagini mediche con alberi decisionali. 2011.

[50] Dewan Md.Fraid, Mohammad Zahidur Rahman, Chowdhury Mofizur Rahman. An Ensemble Approach to Classifier Construction based on Bootstrap Aggregation. International Journal of Computer Application / 0975-8887). Volume 25-N0.5, July 2011.

[51] Robert Bryll, Ricardo Gutierrez-Osuna, Francis Quek. Attribute bagging: improving accuracy of classifier ensembles by using random feature subsets. Pattern Recognition Society, 2002.

[52] Jin Wang, Bo Yun, Pingli Huang, and Yu-Ao Liu. Applying Thresold SMOTE Algorith with Attribute Bagging to Imbalanced Datasets. P.Lingras et al (Eds): RSKT, pp 221-228, 2013.

[53] Hui Zhu, Siyu Chen, Lexiang Zhu, Hui Li, Xiaofeng Chen. RangeTree: A feature Selection Algorith for C 4.5 Decision Tree. 2013.

[54] http://en.wikipedia.org/wiki/C4.5_algorithm;
[55] Marco Masin, Stima della durata del ticket tramite metodi di analisi della sopravvivenza. 2011.
[56] Leo Breiman. Random Forest. 2001
[57] Giacomo Cecchinato. Analisi di terne contigue per il riconoscimento di interfacce. 2013
[58] Ching- Wei Wang, Wun-Hong You. Boosting SVM: effective learning with reduced data dimension. Springer Science+Business Media, 2013.
[59] Michael Pazzani, Pedro Domingos. On the Optimality of the Simple Bayesian Classifier under Zero-One- Loss. Volume 29 Issue 2-3 pp 103-130, November 1997.
[60] Federica Calabretti. Raffronto fra metodi di apprendimento di reti bayesiane su un insieme di dati reali. 2008.

Ringraziamenti

La notte, la luna, le stelle hanno favorito il ricordo e riportato alla mente i volti e i gesti di amici e parenti che mi hanno accompagnato in questo percorso.

Al Professore Dino Pedreschi, che mi ha trasmesso la curiosità e il fascino di queste discipline e per la disponibilità e cortesia dimostratemi.

Al dott. Giulio Rossetti, che in tutti questi mesi mi ha seguito come un'ombra e mi ha fornito le basi e i mezzi per affrontare le problematiche trattate.

Ai miei cugini, Alessandro, Michele, Kety, Sabrina, Daniele, Flavio e Greta, per essere sempre stati presenti in tutte le fasi della mia vita.

A Daniele, amico fraterno, con cui ho condiviso tante avventure e molte risate (quante ne abbiamo combinate) e che è sempre stato presente nei momenti difficili.

A Elena e Chiara, straordinarie amiche e colleghe, che hanno sopportato la mia ingestibilità e sana follia.

A Marzia, la “segretaria bolognese”, che con quel suo non chetarsi mai mi ha fatto venire le orecchie come l'elefantino Dumbo, ma che mi ha permesso, pure, di arricchirmi umanamente.

A Luisa, l'efficientissima segretaria tutto fare e l'essenza dell'amicizia vera, basata sulla condivisione di ideali e valori.

E infine , anche se non sono qui riportati, tutte quelle persone che mi hanno fatto emozionare e sorridere.