Link Prediction su Reti Multidimensionali

Giulio Rossetti Michele Berlingerio Fosca Giannotti
KDDLab, ISTI-CNR, Pisa, Italy
{name.surname}@isti.cnr.it

Sommario L'analisi di reti complesse è un campo di ricerca interdisciplinare, che vede coinvolti fisici, sociologi, matematici, economisti e informatici. In questo articolo estendiamo la formulazione classica del problema del Link Prediction allo scenario delle reti multidimensionali, ossia quelle reti che ammettono più di un link fra due entità. Introduciamo una nuova formulazione che tenga conto delle informazioni multidimensionali espresse dalle reti analizzate, e alcune famiglie di predittori progettati appositamente per sfruttare tali informazioni. Presentiamo infine una valutazione sperimentale dell'applicazione delle soluzioni proposte a reti multidimensionali reali. I risultati preliminari ottenuti sono incoraggianti, e spingono verso una ricerca più estensiva di soluzioni al problema del Link Prediction su reti multidimensionali.

1 Introduzione

Durante gli ultimi anni si è assistito al sorgere, nella comunità scientifica, di un grande interesse per l'analisi e l'estrazione di conoscenza dalle reti complesse che oggi dominano il mondo reale. Una delle direzioni di ricerca è legata all'analisi e la comprensione delle dinamiche evolutive che regolano nel tempo la struttura delle reti. Il tempo nelle reti può giocare un duplice ruolo: nel primo la struttura della rete evolve e si assiste all'apparizione di nuovi nodi o archi, mentre nel secondo, al passare del tempo, vengono compiute delle azioni da parte dei nodi (gli utenti di un social network si scambiano messaggi, ricercatori collaborano alla stesura di articoli, e via discorrendo). A causa della dinamicità delle reti, è comprensibile che l'interesse di numerosi lavori si sia incentrato sulla ricerca di pattern evolutivi che siano in grado di predire come queste evolvano nel tempo.

Recentemente, alcuni ricercatori si sono accorti che molte delle reti reali sono *multidimensionali*, ossia una coppia di nodi può essere connessa da più relazioni, che chiamiamo *dimensioni*. Differenti dimensioni possono rappresentare o tipi di relazione differenti (amicizia, parentela, ecc.), oppure diversi valori dello stesso tipo di relazione (come per esempio la collaborazione in diverse conferenze). Nel mondo reale esistono diversi esempi di reti multidimensionali, come la rete completa dei trasporti (dove treno, autobus, aereo, e nave sono quattro delle dimensioni possibili), le reti sociali (dove due persone possono essere connesse perché sono amiche, o giocano nella stessa squadra, o partecipano agli stessi eventi), o le reti di collaborazione (dove ogni conferenza è una dimensione diversa). Appare chiaro come questo tipo di reti, rispetto a quelle usualmente studiate in letteratura, presentino un ulteriore grado di libertà nella loro complessità, ossia

le dimensioni. Diventa infatti interessante studiare le relazioni che intercorrono fra diverse dimensioni, l'importanza di una dimensione sulle altre, il fatto che due dimensioni si escludano a vicenda, e via discorrendo. In questo scenario, pensando al problema dell'analisi dell'evoluzione delle reti, una possibile domanda interessante da porsi è in quale o quali dimensioni apparirà più probabilmente un nuovo arco.

La multidimensionalità comporta la necessità di sostituire il modello usualmente utilizzato come base per l'analisi di reti, i grafi semplici, con uno più espressivo, i multigrafi. Come conseguenza di tale sostituzione, diviene necessario rivedere gli approcci algoritmici già studiati per i principali problemi di Data Mining in modo da tener conto delle ulteriori informazioni topologiche a disposizione per l'analisi. Congiuntamente alle informazioni inerenti la dimensionalità è interessante, soprattutto per gli studi incentrati sull'evoluzione delle reti, avere la possibilità di osservare, con il maggior dettaglio possibile, la cronistoria della rete oggetto di analisi. Una descrizione dettagliata della storia evolutiva di una rete consente, infatti, di raffinare i risultati ottenuti sfruttando una ancora più vasta fonte di informazioni.

In questo articolo ci occupiamo dell'evoluzione temporale di una rete multidimensionale introducendo un'estensione multidimensionale del problema del Link Prediction (la predizione di nuovi archi in una rete in evoluzione), e definiamo formalmente alcune classi di predittori che tengano conto delle informazioni di correlazione e anticorrelazione tra le dimensioni. Presentiamo quindi, brevemente, alcune reti multidimensionali, utilizzate per l'analisi sperimentale, e illustriamo i principali risultati preliminari ottenuti.

2 Definizione del problema

In [7], il problema del Link Prediction viene definito, data una rete sociale osservata ad un istante temporale t, come il problema di predire gli archi che si andranno ad aggiungere ad essa negli istanti successivi a t. Una volta introdotta la multidimensionalità tale formulazione del problema deve essere necessariamente estesa: le nuove informazioni fornite dalla rete introducono una maggiore complessità poiché i risultati proposti dai modelli predittivi devono riuscire a discriminare le dimensioni in cui compariranno gli archi predetti. Diamo di seguito una definizione formale del problema per l'ambito multidimensionale.

Definizione 1 (Link Prediction Multidimensionale) *Sia $G = (V, E, L, T)$ un multigrafo, non orientato, definito dagli insiemi finiti V dei nodi, E degli archi, L delle dimensioni e T degli istanti temporali associati agli archi.*

Il problema del Link Prediction, dato un multigrafo G ed un istante temporale $t' > max\{t : t \in T\}$ ha come obiettivo predire gli archi che entreranno a far parte del grafo originario in tale istante futuro e la specifica dimensione in cui essi si formeranno.

Per ogni arco predetto - identificato dalla tripla $(nodo, nodo, dimensione)$ - deve inoltre essere calcolato uno score *di confidenza della predizione.*

Come vediamo, quindi, la multidimensionalità aggiunge un grado di libertà al problema: non ci interessa solo sapere quali coppie di nodi si connetteranno in futuro, ma vogliamo anche sapere in quali dimensioni questo avverrà. In sezione 4 introduciamo alcune misure multidimensionali a supporto della soluzione al problema, mentre in sezione 5 introduciamo alcune famiglie di predittori.

3 Lavori Correlati

Molte pubblicazioni hanno trattato il problema del Link Prediction in reti monodimensionali e i principali modelli evolutivi sfruttati per effettuare un'analisi predittiva sulle reti sociali. Le soluzioni predittive analizzate in letteratura possono essere suddivise in *non supervisionate* e *supervisionate* a seconda che queste propongano formule chiuse, indipendenti dalla rete da analizzare, per effettuare la predizione, oppure guidino la fase predittiva tramite l'ausilio di informazioni topologiche estratte dalla rete stessa.

Fra i predittori non supervisionati, in [8] viene introdotta una soluzione basata sul principio del preferential attachment, mentre in [1] viene introdotto un modello basato invece sulle caratteristiche quantitative dei nodi comuni. Una buona rassegna dei modelli non supervisionati è [7], in cui gli autori confrontano empiricamente molti dei classici approcci di questo tipo. Nel nostro articolo, modifichiamo alcuni di questi predittori in modo da adattarli allo scenario multidimensionale.

Due predittori supervisionati sono invece [5] e [4], il primo dei quali, basato sull'estrazione di regole evolutive frequenti [2], non solo è anche in grado di predire nuovi nodi, ma utilizza tutta la storia temporale degli archi presenti nella rete.

In [6] gli autori presentano il problema di predire archi positivi (trust) e negativi (distrust). Sebbene questo possa sembrare una formulazione multidimensionale del Link Prediction, il problema affrontato è in realtà di classificazione, poichè viene predetta solo l'etichetta dell'arco.

4 Misure multidimensionali

Come abbiamo visto, nella letteratura del Link Prediction, molti predittori si basano su misure strutturali calcolate sulla rete. In analogia, i modelli che proponiamo in questa sezione sono basati su misure multidimensionali. Innanzitutto, quindi, vediamo qui alcune misure su reti multidimensionali su cui i nostri predittori sono basati. Per la natura preliminare di questo lavoro, presentiamo qui solo alcune delle misure utilizzate.

L'introduzione della multidimensionalità nel modello analizzato causa la necessità di rivedere parte delle misure utilizzate comunemente nell'analisi di reti: in particolare è necessario definire nuovamente i concetti di *neighbors* e di *degree* di un nodo. Mentre, infatti, in un grafo semplice non orientato, le due coincidono, ciò non accade più in un multigrafo.

Facendo riferimento a [3], in cui gli autori hanno introdotto un framework per l'analisi di reti multidimensionali, riportiamo una variante multidimensionale della misura di *neighbors*, una misura atta a cogliere l'importanza che una specifica dimensione può avere nella rete sulle altre, e due misure che contano la frazione di nodi e di archi presenti in una dimensione. Introduciamo, infine, una nuova misura multidimensionale, che estende il framework citato, mirata a pesare un arco in una dimensione relativamente alla sua storia temporale.

Definizione 2 (Neighbors$_{Xor}$) *Sia $v \in V$ un nodo di una rete G e $D \in L$ un insieme di dimensioni. La funzione $Neighbors_{Xor}$: $V \times P(L) \to N$ (dove $P(L)$ identifica l'insieme delle parti di L), definita come:*

$$Neighbors_{Xor}(v, D) = \sum_{u \in V} k_{uv}(D) \tag{1}$$

dove

$$k_{uv}(D) = \begin{cases} 1 \; se \; \forall (u, v, d) \in E : d \in D \\ 0 \qquad\qquad\qquad altrimenti \end{cases} \tag{2}$$

calcola il numero di vicini del nodo v raggiungibili tramite dimensioni in D, e non tramite archi etichettati con altre dimensioni.

Definizione 3 (Dimension Relevance Weighted) *Sia $v \in V$ un nodo in una rete G e $D \in L$ un insieme di dimensioni. La funzione Dimension Relevance Weighted: $V \times P(L) \to [0, 1]$, è definita come:*

$$DR_W(v, D) = \frac{\sum_{u \in N} Neighbors_{Set}(v, D) \frac{n_{uvd}}{n_{uv}}}{Neighbors(v, L)} \tag{3}$$

dove: $Neighbors_{Set}(v, D)$ identifica il numero dei nodi vici a v raggiungibili tramite archi etichettati da dimensioni appaetenenti all'insieme D, $Neighbors(v, L)$ identifica il numero dei nodi vicini a v raggiungibili tramite archi appartenenti ad una qualsiasi dimensione, n_{uvd} denota il numero di dimensinoni in cui compaiono archi tra i nodi u e v appartenenti a D, e n_{uv} denota il numero di dimensioni in cui compaiono archi tra u e v. Questa misura calcola la frazione di vicini direttamente raggiungibili dal nodo v seguendo archi appartenenti solo alle dimensioni incluse in D, pesata rispetto alle altre possibili dimensioni connettenti ciasun vicino.

Definizione 4 (Ndd) *Dati V l'insieme dei vertici, D l'insieme delle dimensioni (con $d \in D$) e E l'insieme degli archi della rete si definisce il coefficiente di Node Dimension Degree come:*

$$Ndd(d) = \frac{|\{u \in V | \exists v \in V : (u, v, d) \in E\}|}{|V|} \tag{4}$$

Definizione 5 (Edd) *Dati V l'insieme dei vertici, D l'insieme delle dimensioni (con $d \in D$) e E l'insieme degli archi della rete si definisce il coefficiente di Edge Dimension Degree come:*

$$Edd(d) = \frac{|\{(u, v, d) \in E | u, v \in V\}|}{|E|} \tag{5}$$

Definizione 6 (*Wpres*) *La funzione Wpres calcola il totale degli istanti - pesato in base all'ordine temporale - in cui un determinato arco è comparso nella rete nella dimensione specificata. Archi comparsi in istanti recenti hanno un peso maggiore.*

$$Wpres(u, v, d) = \sum_{\{t|(u,v,d,t)\in E\}} \Pi_t \tag{6}$$

dove Π_t indica il peso dell'istante temporale t.

5 Modelli Predittivi

Per affrontare il problema del Link Prediction sono stati proposti, in letteratura, molteplici modelli sia supervisionati che non supervisionati. La grande vastità e diversità dei modelli definiti deriva dall'aver ipotizzato, o estratto, differenti pattern evolutivi da eterogenee tipologie reti. In analogia, in questa sezione, introduciamo diverse classi di predittori da noi definiti. Per la natura preliminare di questo lavoro e per ragioni di spazio, introduciamo solo due famiglie di predittori. Facciamo notare, però, che è possibile combinare diversi predittori fra loro, oppure moltiplicarli per una o più delle misure sopra definite, in modo da catturare più segnali sovrapposti nell'evoluzione della rete. Questo è esattamente l'approcio seguito in [5], dove gli autori presentano diverse combinazioni fra il loro predittore basato sulle regole di evoluzione frequenti e alcuni predittori classici come Adamic-Adar e Common Neighbors.

Come prima soluzione, abbiamo modificato alcuni dei classici modelli non supervisionati (sono stati presi in considerazione Adamic-Adar, Common Neighbors e Jaccard) e li abbiamo adattati allo scenario multidimensionale. Per fare ciò, abbiamo semplicemente sostituito la misura $Neighbors$ da essi utilizzata, con la sua variante $Neighbors_{Xor}$ in modo da poter discriminare la dimensione degli archi predetti. In questo modo abbiamo creato una base sperimentale per il confronto dei successivi predittori, le cui performance sono state valutate rispetto a questi modelli.

Una soluzione più avanzata è stata quella di combinare i tre predittori base con le misure multidimensionali Ndd, Edd, e $Wpres$ (utilizzate come coefficienti moltiplicativi degli score) in modo da poter valutare se l'introduzione di ulteriori informazioni sulla struttura multidimensionale della rete riescono a garantire predizioni migliori.

L'ultima soluzione sperimentata è basata sulla misura DR_W introdotta in precedenza. Contrariamente a quanto fatto finora, non utilizziamo i modelli base nella predizione, ma ci affidiamo esclusivamente ad un modello ideato ad hoc e alle informazioni temporali fornite dalla rete.

Definizione 7 (*WDR*) *Siano $u, v \in V$ nodi della rete e $d \in D$ una dimensione: il predittore WeightedDimensionRelevance con informazioni di Weighted Presence è definito come:*

$$WDR(u, v, d) = DR_W(u, d) * DR_W(v, d) * (1 + Wpres(u, v, d)) \tag{7}$$

Durante la fase sperimentale abbiamo in realtà definito e provato diverse combinazioni di predittori, aggregati (non solo moltiplicazione), e coefficienti. Tuttavia, per ragioni di spazio e data la natura preliminare del lavoro, vediamo solo gli esperimenti relativi ai predittori sopra definiti.

6 Analisi Sperimentale

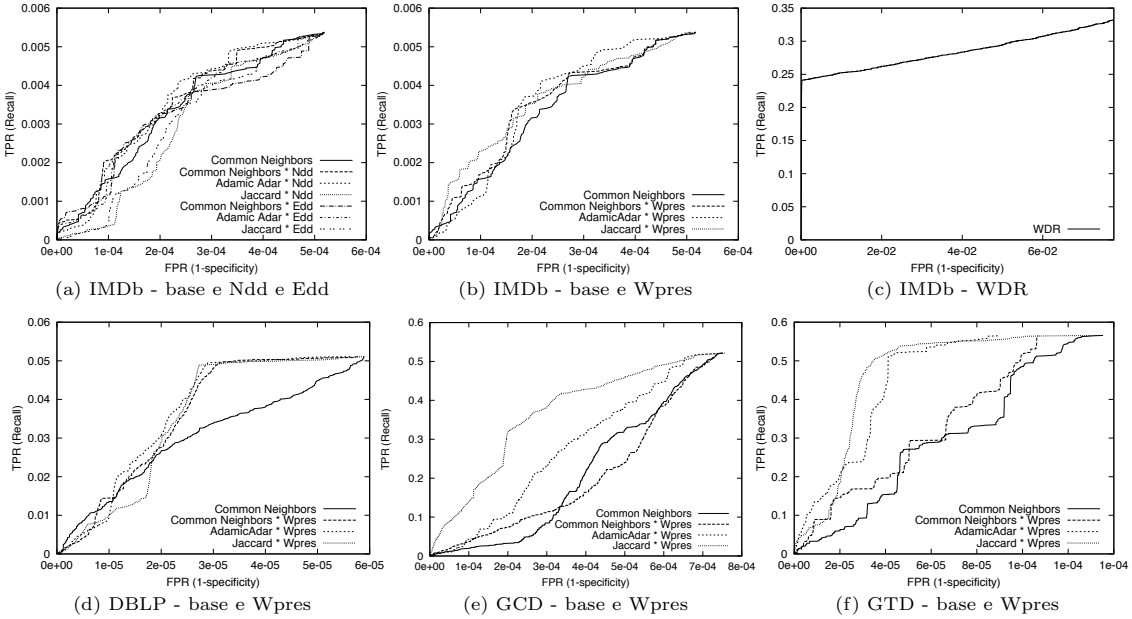
In questa sezione vediamo i risultati preliminari ottenuti applicando i nostri predittori ad alcune reti reali. Ciasuna rete è stata divisa in *training* e *test* set, e l'accuratezza dei predittori è stata misurata sul test set tramite curve ROC. Tale modalità di comparazione è stata scelta poiché consente una più facile lettura dei risultati ottenuti rispetto a quanto offerto dalle curve di Precision/Recall.

6.1 Reti Multidimensionali

Abbiamo utilizzato reti di diversa natura, ed in particolare:

- IMDb[1]: rete di attori cinematografici. Due attori sono connessi se hanno partecipato ad uno stesso film in uno stesso anno, e le dimensioni rappresentano la tipologia del film. Nodi: 43.867, dimensioni: 28. Il training set è composto da 216.544 archi appartenenti ad un arco temporale di 10 anni (1998-2007), il test set da 54.749 archi appartenenti all'anno successivo.
- DBLP[2]: rete di collaborazioni scientifiche. Due autori sono connessi se hanno partecipato alla stesura di un articolo in uno stesso anno, e le dimensioni sono definite dalle specifiche conferenze. Nodi: 30.176, dimensioni: 28. Il training set è composto da 78.956 archi appartenenti ad un arco temporale di 10 anni (1998-2007), il test set da 14.650 archi appartenenti all'anno successivo.
- GCD[3]: rete di fumettisti. Due fumettisti sono connessi se hanno partecipato alla stesura di uno stesso numero di un fumetto in un determinato arco temporale. Le dimensioni rappresentano la sezione del fumetto a cui hanno collaborato (ad es. copertina, storia). Nodi: 10.000, dimensioni: 6. Il training set è composto da 140.546 archi appartenenti all'arco temporale 1990-1999, il test set da 4.945 archi appartenenti all'anno successivo.
- GTD[4]: rete di gruppi terroristici. Due gruppi terroristici sono connessi se hanno pianificato un attentato nello stesso arco temporale in uno stesso stato. Gli stati rappresentano le dimensioni. Nodi: 2.755, dimensioni: 209. Il training set è composto da 25.200 archi, appartenenti all'arco temporale 1970-2007, il test set 2.572 archi appartenenti all'anno successivo.

[1] http://www.imdb.com
[2] http://dblp.uni-trier.de
[3] http://www.comics.org
[4] http://www.start.umd.edu/gtd



(a) IMDb - base e Ndd e Edd (b) IMDb - base e Wpres (c) IMDb - WDR

(d) DBLP - base e Wpres (e) GCD - base e Wpres (f) GTD - base e Wpres

Figura 1. Alcuni dei risultati ottenuti. Sull'asse delle ascisse è riportato i valori di FPR, mentre sulle ordinate i valori di TPR.

6.2 Risultati

In Figura 1 riportiamo le curve ROC calcolate sui risultati ottenuti sulle nostre reti. La prima riga mostra le differenze fra le varie famiglie di predittori sulla rete IMDb. In Figura 1(a) mostriamo l'andamento dei predittori di base moltiplicati per i coefficienti multidimensionali globali: utilizzando Common Neighbors come modello base di riferimento è possibile notare come gli incrementi nelle performance risultino contenuti seppur presenti. In Figura 1(b) abbiamo i predittori base estesi per l'analisi temporale tramite il moltiplicatore Wpres. In questo caso gli incrementi delle performance predittive risultano più sensibili. La Figura 1(c) riporta l'andamento del predittore WDR: in questo grafico si è omesso il confronto con i modelli base a causa della diversa scala utilizzata per la definizione del grafico. La seconda riga riporta invece i predittori base moltiplicati con Wpres, nelle altre tre reti. Come si può osservare, in queste reti la conoscenza multidimensionale e la storia temporale completa degli archi favorisce notevolmente l'accuratezza della predizione.

Un'analisi generale dei risultati suggerisce come l'andamento delle performance sia strettamente legato alla reale topologia della rete in esame. Questo risultato, per quanto non consenta la determinazione di un predittore *universale*, ossia di un modello con performance sempre superiori indipendentemente dalla rete analizzata, era attendibile data la tipologia dei modelli introdotti (non supervisionati basati su neighbors).

Quanto a Edge/Node Dimension Degree e Weighted Presence, sulle reti analizzate possiamo affermare che l'introduzione di moltiplicatori multidimensionali

globali alla rete, e temporali locali ai singoli archi, consente di incrementare le performance dei modelli base rispettivamente nel 40-60%, per Ndd e Edd, e nel 60%, per $Wpres$, dei test eseguiti. Tale incremento varia a seconda del modello base a cui i coefficienti moltiplicativi sono applicati.

Tramite l'adozione del modello ad hoc WDR è possibile ottenere, per le reti analizzate, i migliori risultati sia per numero di predizioni corrette, sia per l'assegnazione degli score di confidenza ad ogni predizione.

7 Conclusioni e Lavori Futuri

Abbiamo presentato un'estensione del problema del Link Prediction allo scenario delle reti multidimensionali. Guidati da diverse misure multidimensionali topologiche e storiche, abbiamo definito diverse famiglie di predittori per queste reti. I risultati preliminari ottenuti sono incoraggianti, e sembrano confermare la validità dell'approccio, e l'effettiva potenza delle misure multidimensionali nel catturare fenomeni legati anche al modello evolutivo delle reti.

In futuro, ci proponiamo di estendere l'analisi sperimentale al fine di caratterizzare le reti secondo la loro prevedibilità temporale, tramite le misure e i predittori proposti. Inoltre, intendiamo verificare la fattibilità dell'introduzione di modelli supervisionati per risolvere il problema del Link Prediction su reti multidimensionali.

Riferimenti bibliografici

1. Lada A. Adamic and Eytan Adar. Friends and neighbors on the web. *Social Networks*, 25(3):211–230, 2003.
2. Michele Berlingerio, Francesco Bonchi, Björn Bringmann, and Aristides Gionis. Mining graph evolution rules. ECML PKDD '09, pages 115–130, Berlin, Heidelberg, 2009. Springer-Verlag.
3. Michele Berlingerio, Michele Coscia, Fosca Giannotti, Anna Monreale, and Dino Pedreschi. Foundations of multidimensional network analysis. Tech. Rep. 2010-TR-004. http://puma.isti.cnr.it/dfdownload.php?ident=/cnr.isti/2010-TR-004, 2010.
4. Mustafa Bilgic, Galileo Mark Namata, and Lise Getoor. Combining collective classification and link prediction. ICDMW '07, pages 381–386, Washington, DC, USA, 2007. IEEE Computer Society.
5. Björn Bringmann, Michele Berlingerio, Francesco Bonchi, and Arisitdes Gionis. Learning and predicting the evolution of social networks. *IEEE Intelligent Systems*, 25(4):26–35, 2010.
6. Jure Leskovec, Daniel P. Huttenlocher, and Jon M. Kleinberg. Predicting positive and negative links in online social networks. In Michael Rappa, Paul Jones, Juliana Freire, and Soumen Chakrabarti, editors, *WWW*, pages 641–650. ACM, 2010.
7. David Liben-Nowell and Jon Kleinberg. The link prediction problem for social networks. CIKM '03, pages 556–559, New York, NY, USA, 2003. ACM.
8. 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. *PNAS*, 99(8):5207–5211, April 2002.