UNIVERSITÀ DEGLI STUDI DI BARI ALDO MORO

DIPARTIMENTO DI INFORMATICA

Dottorato di ricerca in Informatica e Matematica
XXXII ciclo

Progetto di ricerca

Dottorando: Dott. Paolo Mignone

Tutor: Prof. Michelangelo Ceci

Firma del dottorando _____

Firma del tutor _____

1. Titolo della ricerca:

Link prediction via transfer learning across multiple heterogeneous domains

2. Area nella quale si inquadra la ricerca:

Apprendimento automatico e Data Mining

3. Motivazioni e obiettivi della ricerca

Link prediction è uno specifico task di data mining che sfrutta la modellazione relazionale dei dati. Nel link prediction l'obiettivo è quello di stimare la probabilità di una relazione tra due entità, di cui allo stato attuale non vi è alcuna informazione.
La modellazione relazionale dei dati ha continuato ad attrarre sempre più interessi e negli ultimi anni ha trovato applicazioni di successo in molti settori quali l'analisi di social network, delle reti di geolocalizzazione, dei recommender system e altri in cui i dati si dispongono naturalmente su una struttura a rete.
Un'assunzione di base degli algoritmi di machine learning classici è quella di disporre di training e testing set provenienti dalla stessa distribuzione di dati, dallo stesso dominio e descritti nello stesso spazio delle feature. Tuttavia, in molti contesti reali, questa assunzione è troppo forte. Infatti, nello studio delle reti biologiche, sociali e tecnologiche, collezionare i dati di training è molto dispendioso. Da qui l'esigenza di costruire modelli basati su dati già disponibili in contesti diversi, ma correlati. Strategie di *transfer learning* consentono di sfruttare la conoscenza proveniente da un task sorgente per migliorare le performance di un task target per il quale è presente una scarsa quantità di dati.

Obiettivo della ricerca sarà pertanto focalizzato sulla progettazione e implementazione di metodi per il link prediction, applicabili anche a domini con scarsa disponibilità di dati etichettati, sfruttando le emergenti strategie di transfer learning. Pertanto, si affronteranno problemi noti riguardanti il link prediction come: 1) aggiornamento del modello a seguito di aggiunta o rimozione di nodi e archi della rete; 2) comprensione del cambiamento nel tempo dei pattern di interazione; 3) comprensione di come l'interazione tra due nodi può essere influenzata da altri nodi o altre interazioni; 4) gestione delle reti con alto livello di sparsità e sbilanciamento di classe dei dati. Inoltre verranno affrontati i problemi riguardanti il transfer learning come: 1) evitare il negative transfer, che avviene quando la strategia di transfer learning porta ad un peggioramento delle performance dei modelli risultanti; 2) definire soluzioni generali che indichino il livello di

trasferibilità di conoscenza tra due o più distinti domini e task; 3) utilizzare solo le parti di un dominio utile al trasferimento di conoscenza. Individuando le soluzioni a questi problemi, si potrebbero raggiungere traguardi molto interessanti in vari contesti scientifici (come quello biologico, medico, astronomico) e industriali (per l'analisi dei social network, il marketing e l'implementazione di recommender system collaborativi).

4. Stato dell'arte

Gli algoritmi di transfer learning possono essere classificati in **omogeneo** ed **eterogeneo**. Al fine di chiarirne le differenze, è utile introdurre il concetto di dominio: un dominio D è definito come
D = {*X*, P(X)}, dove *X* rappresenta lo spazio delle feature, P(X) la distribuzione di probabilità marginale dei dati, con $X = \{x_1, x_2, \dots, x_n\} \in X$. Considerando D_S e D_T rispettivamente dominio sorgente e target, ha senso parlare di transfer learning quando $D_S \neq D_T$.
Coerentemente con questa definizione, pertanto, gli algoritmi di transfer learning omogenei lavorano su diverse distribuzioni di probabilità marginali dei dati di training e stesso spazio delle feature, mentre gli algoritmi di transfer learning eterogenei lavorano su diversi spazi delle feature.
Il survey di Weiss [12] fornisce una panoramica particolarmente significativa sugli studi e sulle nuove applicazioni del transfer learning.
I lavori in cui si affronta il task di link prediction mediante strategie di transfer learning sono limitati. Tra questi c'è [1], in cui gli autori propongono un modello Bayesiano non parametrico che considera la somiglianza tra diversi task sfruttando tutti i dati di collegamento insieme. I risultati mostrano che la strategia di transfer learning supporta positivamente l'addestramento dei modelli predittivi. Tuttavia, non è specificato il criterio di scelta dei domini da cui effettuare trasferimento di conoscenza e in che misura essi possano esprimere la trasferibilità della loro conoscenza.
Uno degli ambiti scientifici in cui sono stati applicati metodi di link prediction è quello biologico, in particolare per l'analisi di dati di espressione genica. Sempre in campo biologico, in lavori più recenti è stata proposta l'applicazione di tecniche di link prediction anche a interazioni tra microRNA (miRNA) e messenger RNA (mRNA) [2].
Nell'analisi dei social network, è più efficace utilizzare le feature della struttura a rete piuttosto che le informazioni dei singoli nodi, al fine di effettuare con più precisione le previsioni sulle loro interazioni [3].
In [4] viene presentata una semplice ma efficace strategia di previsione di interazione basata sulla propagazione di etichette nella rete, imitando la comunicazione tra le persone. Inoltre, è possibile effettuare link prediction anche in ambiti in cui le caratteristiche della rete variano nel tempo, come avviene in [5].

In un recente lavoro [6], viene prevista l'insorgenza di sintomi futuri di alcuni pazienti, sulla base del loro stato di salute attuale. A tal proposito, gli autori prima costruiscono una rete pesata di parametri medici anormali, considerando le interazioni (link) tra tali parametri. Inoltre, viene proposto un metodo di link prediction non supervisionato per identificare le connessioni tra parametri anormali. Infine, gli autori costruiscono la struttura della rete dei parametri anormali la quale evolve rispetto all' età dei pazienti.

In [7], metodi di collaborative filtering supportano la previsione e la raccomandazione di coppie di utenti anche senza alcuna interazione pregressa di fondo. A questo proposito, viene presentato il concetto di "collaborative path": il percorso in comune tra due o più utenti nella rete delle loro attività scelto secondo tre specifiche misure di prossimità.

Anche in ambito linguistico gli autori di [8] propongono un algoritmo di link prediction tra coppie di documenti di lingue differenti ed effettuano una previsione sulle parole di un testo a prescindere dalla lingua.

Tuttavia, in letteratura, gli studi che effettuano task di link prediction con dati non sufficientemente etichettati sono limitati, sebbene tale situazione si verifichi in molti contesti reali.

Molti studi recenti dimostrano che è possibile lavorare in domini con pochi dati etichettati a disposizione sfruttando modelli addestrati in altri domini con maggiori disponibilità. In uno studio molto interessante [9] viene effettuata una caratterizzazione della fatica muscolare sui dati ottenuti attraverso un metodo di rilevamento non invasivo come l'elettromiografia superficiale. L'autore segue una strategia di transfer learning omogeneo: i dati dei domini sorgente (di cui si dispone una grande mole di dati etichettati) presentano lo stesso spazio delle feature del dominio target (di cui non si dispone di dati etichettati a sufficienza). L'idea di base è quella di utilizzare una combinazione dei classificatori costruiti sui domini sorgente per etichettare le osservazioni non etichettate del dominio target mediante delle "pseudo-etichette". Successivamente costruisce un modello predittivo sui dati etichettati e pseudo-etichettati nel task target.

Un altro lavoro interessante è quello proposto in [10], dove gli autori propongono di costruire un modello predittivo sulle immagini di Flickr e Google immagini (task sorgente) per poi applicarlo per classificare eventi nei video (task target): il modello appreso nel task sorgente viene impiegato per classificare frame chiave dei video.

In un altro lavoro [11], l'obiettivo è quello di ridurre la sparsità delle matrici dei recommender system collaborativi. In particolare, gli autori propongono di costruire un modello predittivo su dati di rating binari (like/dislike) per poi applicarlo a dati relativi a ratings non binari (scala da 1 a 5).

5. Approccio al problema

L'approccio al problema sarà incentrato sulla risoluzione delle problematiche presentate precedentemente:

- tecniche di transfer learning omogenee richedono solitamente meno effort, pertanto, laddove sia possibile, potrebbe essere utile omogenizzare preventivamente i due domini di interesse al fine di applicare strategie più economiche. In letteratura, questa soluzione è chiamata *domain adaptation*. Tuttavia, in molti casi potrebbe essere molto complesso o più dispendioso rimodellare i domini piuttosto che procedere con una strategia di transfer learning eterogeneo;
- per aggiornare il modello predittivo a seguito di una modifica della rete è possibile procedere in maniera statica, effettuando nuovamente la fase di addestramento del modello considerando le nuove informazioni, o in maniera dinamica, in cui è prevista una fase di addestramento incrementale che prevede l'uso di tecniche specifiche, quali le *sliding windows*, che tengano conto del fattore tempo e del rapido cambiamento dei dati;
- il problema dello sbilanciamento di classe dei dati, molto comune in ambiti biologici, può essere risolto mediante tecniche di *ensemble learning*;
- saranno proposte delle metriche che tengano conto delle distanze, sia tra gli spazi delle feature, che tra le distribuzioni di probabilità marginali dei dati nei due domini di interesse, al fine di quantificare il livello di trasferibilità di conoscenza. Tramite tali metriche si potrebbero prevenire eventuali problematiche di negative transfer e supportare la costruzione di un framework teorico.

Inoltre, per progettare metodi di link prediction utili in contesti reali, sarà necessario procedere per fasi:

1) acquisire dei dati reali dello specifico contesto scientifico o industriale;
2) organizzare tali dati in una struttura a rete: solitamente, i dati in fase di acquisizione non sono disponibili sotto forma di relazioni (ad esempio i dati genici sono espressi inizialmente da singoli geni e le amicizie sui social network si riferiscono a singoli individui);
3) costruire modelli descrittivi per ottenere informazioni topologiche globali e locali delle reti di interazione;
4) implementare modelli predittivi dell'esistenza delle interazioni anche associate ad un valore che ne esprima la forza, la probabilità, l'affidabilità;
5) gestire, mediante strategie di transfer learning, problematiche note quali il controllo della sparsità della rete, dovuta alla scarsa quantità di collegamenti etichettati.

6. Risultati attesi

L'uso di tecniche di link prediction può essere utile in numerosi ambiti applicativi:

1) medicina e biologia, per l'individuazione di sottoreti di geni che presentano caratteristiche funzionali ricorrenti e indipendenti dallo specifico organismo (networks motifs); per lo studio di farmaci che presentano effetti simili; per lo studio dei dati di espressione genica in pazienti sani e malati, evidenziando le eventuali cause che portano le cellule a livelli di espressione genica cancerogena;
2) social network analysis, per l'individuazione di possibili interazioni tra persone non ancora in relazione; per la raccomandazione di contenuti ad un singolo individuo sfruttando le sue informazioni di interazione;
3) e-commerce, per supportare la costruzione di recommender system che possano suggerire prodotti che soddisfino le esigenze degli utenti;

Inoltre, lo studio di strategie di transfer learning sarà ortogonale allo studio del task di link prediction: il trasferimento di conoscenza tra diversi domini applicativi è una emergente strategia che consente di costruire modelli descrittivi e predittivi con notevole risparmio di tempo ed effort per la raccolta dei dati, non solo per il link prediction. Pertanto è un topic trasversale a contesti sia scientifici che industriali.

7. Fasi del progetto

Anno 1°: studio della letteratura, dello stato dell'arte e del materiale di ricerca di base:

Attività 1.1: studio approfondito di aspetti teorico-formali dell'apprendimento automatico e dello sviluppo di sistemi per la scoperta di conoscenza dai dati;

Attività 1.2: approfondimento delle tematiche relative alle tecniche di classificazione applicate a dati rappresentati tramite reti omogenee ed eterogenee;

Attività 1.3: ricerca e studio di metodi di trasferimento di conoscenza da uno o più domini sorgenti verso uno specifico dominio target in accordo con gli obiettivi di ricerca;

Attività 1.4: partecipazione a scuole internazionali e conferenze su argomenti inerenti all'attività e agli obiettivi previsti.

Anno 2°: sintesi, realizzazione e implementazione di metodi:

Attività 2.1: studio delle attività svolte da gruppi di ricerca con obiettivi affini;

Attività 2.2: sintesi, progettazione e implementazione di metodi per il link prediction tra oggetti di tipi omogenei ed eterogenei, con e senza strategie di transfer learning;

Attività 2.3: valutazione dei metodi realizzati, confronto con approcci esistenti e pubblicazione dei risultati conseguiti in riviste e/o conferenze internazionali.

Anno 3°: applicazione al dominio applicativo scelto e sviluppo della tesi di dottorato:

Attività 3.1: confronto con l'attività svolta presso altri gruppi di ricerca sia nazionali che internazionali con obiettivi affini mediante eventuali stage presso università o centri di ricerca stranieri;

Attività 3.2: perfezionamento dei metodi con introduzione di accorgimenti specifici per il dominio applicativo scelto;

Attività 3.3: analisi dei risultati sperimentali ottenuti sul particolare dominio applicativo scelto;

Attività 3.4: stesura della tesi di dottorato.

8. Valutazione dei risultati

In letteratura esistono misure di qualità ben consolidate per valutare i risultati ottenuti per il task di link prediction:

- $Precision = \frac{TP}{TP + FP}$
- $Recall = \frac{TP}{TP + FN}$

- $TPR = \frac{TP}{TP + FN}$
- $FPR = \frac{FP}{FP + TN}$
- $FMeasure = 2 * \frac{precision * recall}{precision + recall}$

In cui TP indica true positive, FP false positive, TN true negative e FN false negative. TPR indica il true positive rate e FPR il false positive rate. In alcuni contesti è più opportuno parlare di recall (ad esempio nell'information retrieval).

Tali misure tuttavia sono dipendenti da valori soglia da considerare per calcolare i tassi TP, FP, TN e FN. Misure indipendenti da soglie possono essere considerate per ottenere un indice di qualità più oggettivo. Tra queste si distinguono due principali misure:

- **AUROC**, che rappresenta l'area sotto la curva Receiver Operating Characteristic (ROC). La curva ROC si dispone sul piano cartesiano avente per asse delle ascisse i valori di TPR e sull'asse delle ordinate i valori di FPR. La curva viene tracciata al variare di tutte le possibili soglie applicabili per ottenere TPR e FPR.
- **AUPR**, che rappresenta l'area sotto la curva Precision Recall (PR). La curva PR si dispone sul piano cartesiano avente per asse delle ascisse i valori di Precision e sull'asse delle ordinate i valori di Recall. La curva viene tracciata al variare di tutte le possibili soglie applicabili per ottenere Precision e Recall.

9. Eventuali referenti esterni al Dipartimento

Durante gli studi e la partecipazione a scuole estive inerenti
agli obiettivi prefissati, si spera di poter identificare referenti stranieri, operanti
presso Università della Comunità Europea, le cui collaborazioni saranno certamente utili nel lavoro di ricerca. Attualmente, possibili referenti esterni sono:
• Kurt Driessens - Department of Knowledge Engineering, Maastricht University, The Netherlands
• Sašo Džeroski - Department of Knowledge Technologies, Ljubljana, Slovenia

10. Riferimenti bibliografici

[1] Bin Cao, Nathan Nan Liu, Qiang Yang: Transfer Learning for Collective Link Prediction in Multiple Heterogenous Domains. ICML 2010: pp. 159-166.

[2] Gianvito Pio, Donato Malerba, Domenica D'Elia, Michelangelo Ceci: Integrating microRNA target predictions for the discovery of gene regulatory networks: a semi-supervised ensemble learning approach. BMC Bioinformatics 15(S-1): S4 (2014)

[3] Mohammad Al Hasan, Mohammed J. Zaki: A Survey of Link Prediction in Social Networks. Social Network Data Analytics 2011: pp. 243-275.

[4] Jie Liu, Baomin Xu, Xiang Xu, Tinglin Xin: A link prediction algorithm based on label propagation. J. Comput. Science 16: pp. 43-50. (2016)

[5] Ke-Jia Chen, Yang Chen, Yun Li, Jingyu Han:A supervised link prediction method for dynamic networks. Journal of Intelligent and Fuzzy Systems31(1):
pp. 291-299. (2016)

[6] Buket Kaya, Mustafa Poyraz: Unsupervised link prediction in evolving abnormal medical parameter networks. Int. J. Machine Learning & Cybernetics 7(1): pp. 145-155. (2016)

[7] Amin Shahmohammadi, Ehsan Khadangi, Alireza Bagheri: Presenting new collaborative link prediction methods for activity recommendation in Facebook. Neurocomputing 210: pp. 217-226. (2016)

[8] Yosuke Sakata, Koji Eguchi: Cross-lingual link prediction using multimodal relational topic models. ICIS 2016: pp. 1-8.

[9] Rita Chattopadhyay, Jieping Ye, Sethuraman Panchanathan, Wei Fan, Ian Davidson: Multi-source domain adaptation and its application to early detection of fatigue. KDD 2011: pp. 717-725.

[10] Lixin Duan, Dong Xu, Shih-Fu Chang: Exploiting web images for event recognition in consumer videos: A multiple source domain adaptation approach. CVPR 2012: pp. 1338-1345.

[11] Weike Pan, Nathan Nan Liu, Evan Wei Xiang, Qiang Yang: Transfer Learning to Predict Missing Ratings via Heterogeneous User Feedbacks. IJCAI 2011: pp. 2318-2323.

[12] Karl Weiss, Taghi M. Khoshgoftaar and DingDing Wang: A Survey of Transfer Learning. Springer Open. Journal of Big Data. (2016)