



**Dottorato di ricerca in Informatica e Matematica
XXIX ciclo**

Progetto di ricerca

Dottorando: Dott. Francesco Serafino

Tutor: Prof. Michelangelo Ceci

Coordinatore

Prof. Donato Malerba

Firma del dottorando _____

Firma del tutor _____

1) Titolo della ricerca:

Structured Output Prediction in Network Data

2) Area nella quale si inquadra la ricerca:

Apprendimento automatico e Data Mining.

3) Obiettivi della ricerca:

Il presente progetto di ricerca ha come obiettivo quello di sviluppare metodi di modellazione predittiva in grado di sfruttare enormi insiemi di dati organizzati in reti informative, anche solo parzialmente etichettati, il cui output è organizzato nella forma di una struttura dati complessa.

I dati sono organizzati in una rete informativa, dove ogni nodo rappresenta un oggetto e ogni arco una relazione che sussiste tra gli oggetti in essa coinvolti. La rete può essere omogenea (nel caso più semplice) dove tutti i nodi sono dello stesso tipo, oppure eterogenea, dove gli oggetti sono di tipi differenti. In entrambi i tipi di rete accade il fenomeno di autocorrelazione di cui si deve tener conto. Infatti, mentre il fenomeno della correlazione denota una dipendenza che sussiste tra due variabili differenti di uno stesso oggetto, l'autocorrelazione denota la dipendenza che sussiste tra i valori assunti dalla stessa variabile in oggetti differenti ma in relazione tra loro.

L'approccio di predizione può essere supervisionato, nel caso in cui gli oggetti in gioco siano tutti etichettati, oppure semi-supervisionato nel caso in cui non tutti gli oggetti considerati siano dotati di etichetta. In particolare è di interesse un approccio semi-supervisionato trasduttivo che sia in grado di avere conoscenza a priori in merito agli oggetti della rete etichettati e non, in modo che in fase predittiva sia in grado di etichettarli.

L'output dei sistemi di modellazione predittiva è una struttura dati complessa e quindi il task risultante è quello di Structured Output Prediction (SOP), dove non viene predetta semplicemente una singola etichetta per ogni istanza da classificare. Esempi di SOP sono: la Multi-label classification (MLC), dove ad una istanza vengono associate un insieme di etichette di un particolare dominio; la Hierarchical Multi-Label Classification (HMC), che si basa sulla MLC imponendo un vincolo di gerarchia tra le etichette associate ad una istanza; la Multi-Dimensional Classification che associa ad una istanza un insieme di etichette (organizzate gerarchicamente o meno) appartenenti a domini ortogonali tra loro, dove ogni dimensione rappresenta un dominio particolare nel quale si applica la classificazione. Inoltre, quest'ultima tecnica permette di ottenere una predizione che offre più viste della classe di uno stesso oggetto correlate tra loro, infatti se si considerano delle pagine web da classificare, una pagina potrebbe appartenere alla classe "Guerra" e "Cronaca Nera" secondo la dimensione "Topic", e allo stesso tempo potrebbe appartenere alla categoria "Medio Oriente" secondo la dimensione "Luogo". I domini sono differenti ma sono semanticamente interconnessi tra loro.

Il contributo maggiore che si intende dare in questa direzione consiste nello sviluppo di nuovi approcci di SOP che siano in grado sfruttare tipi differenti di reti informative (omogenee/eterogenee), tramite tecniche di apprendimento supervisionato e semi-supervisionato con differenti tipi di output strutturato.

4) Motivazioni della ricerca

I task di Data Mining di SOP sono già noti in letteratura, ma necessitano di ulteriori approfondimenti. Nuove soluzioni, infatti, sono da ricercare in quanto molti dei metodi di predizione di output strutturato, come ad esempio i metodi di MLC, soffrono di problemi di complessità computazionale che ne limitano la loro applicabilità in contesti reali. All'aumentare del numero di esempi o di etichette, questi metodi diventano inapplicabili.

Inoltre, gli attuali approcci SOP non considerano dati parzialmente etichettati, né possono trattare le connessioni che sussistono tra i dati di training qualora i dati fossero organizzati in una rete informativa.

Ad esempio se si considera il dominio biologico, la rete informativa viene utilizzata per rappresentare le PPI network (Protein to Protein Interaction network) e il task di MLC viene utilizzato per la predizione dell'insieme di funzioni (organizzate gerarchicamente) che svolge un particolare gene. In questo ambito sono già sfruttate le relazioni nelle PPI network ed esistono già in letteratura algoritmi di MLC che sfruttano il fenomeno dell'autocorrelazione, ma c'è ancora da lavorare su metodi semi-supervisionati, in grado di sfruttare la conoscenza derivante dalle reti informative che coinvolgono anche oggetti non etichettati.

Per questi motivi lo studio del problema della classificazione, multi-label e gerarchica nelle PPI network deve essere portato avanti al fine di fornire soluzioni semi-supervisionate che siano in grado anche di relazionarsi con il problema di complessità computazionale.

Riassumendo, gli approcci esistenti di SOP, le rappresentazioni dei dati tramite reti informative e le tecniche supervisionate e semi-supervisionate, sono già note in letteratura, ma la combinazione di questi approcci non è stata ancora affrontata. Metodi che possano trattare efficientemente enormi insiemi di dati, parzialmente etichettati e con output strutturato sono ancora ad uno stato embrionale e il loro sviluppo è un obiettivo audace da perseguire.

5) Stato dell'arte

Lo stato dell'arte nella Structured Output Prediction è proliferante di approcci per quanto riguarda il task di SOP nella multi-label classification. Alcuni approcci trattano, invece, il task leggermente più complicato di hierarchical MLC, dove le possibili etichette sono organizzate in una gerarchia.

Gli approcci di SOP possono essere divisi principalmente in approcci locali e approcci globali. I primi decompongono l'output strutturato in componenti scalari e successivamente utilizzano gli approcci classici di classificazione e regressione per la predizione di ogni componente. I secondi adattano gli approcci classici di apprendimento affinché siano in grado di manipolare l'output strutturato direttamente e apprendono dei modelli che predicano valori completamente strutturati piuttosto che predire le componenti primitive che sono successivamente combinate. Diversi approcci di apprendimento sono stati applicati nel contesto della SOP locale, mentre i metodi adattati per la SOP globale includono approcci di apprendimento basati su distanza, kernel e metodi basati su alberi.

Quasi tutti i metodi di SOP, comunque, hanno serie difficoltà di scalabilità rispetto ad enormi dataset. Un dataset di SOP, infatti, può variare in grandezza in vari modi, tra cui la dimensionalità dell'input (considerando il numero di features in input), lo spazio di output (considerando il numero di etichette nel task MLC) e il numero di istanze. Pochi metodi scalano bene su tutti questi aspetti,

infatti: i metodi basati su kernel possono trattare dataset con features molto numerose, ma all'aumentare degli esempi o all'aumentare della dimensionalità dell'output diventano intrattabili e la loro applicabilità diventa limitata.

I principali domini in cui trovano larga applicazione i metodi di SOP sono il dominio biologico, per quanto riguarda la predizione della funzione dei geni, e il dominio di web document classification dove ad un documento possono essere associate più categorie organizzate gerarchicamente, anche appartenenti a gerarchie concettualmente ortogonali tra di loro.

In ambito biologico diversi approcci di machine learning sono stati proposti recentemente a partire dal lavoro iniziato da Pavlidis et al [1]. Alcuni di questi lavori tengono conto del fenomeno di autocorrelazione. In questo caso, si assume che gli oggetti eventualmente correlati tra loro mediante relazioni (e.g. relazione di vicinanza in ambito spaziale o relazione di interazione tra geni) mostrino maggiore contiguità nei valori di classe di quanto sia mostrato da oggetti non correlati. In un lavoro di Pandey et al [2] può essere trovata una analisi approfondita dei maggiori metodi di MLC in ambito biologico ed in particolar modo riguardanti la predizione della funzione delle proteine utilizzando vari tipi di dati biomedici. Da questo lavoro si può affermare che le tecniche migliori da applicare a dati biologici hanno portato alla luce molte sfide, alcune delle quali non ancora risolte. In particolare, l'identificazione dei sottoinsiemi più rilevanti dei dati rispetto alle classi funzionali è ancora un problema aperto, inoltre i dati biologici sono soggetti a rumore ed incompletezze.

La maggioranza dei sistemi che affrontano, invece, il task MLC gerarchica si basano sui Predictive Clustering Tree (PCT). In questi sistemi, gli alberi di decisione sono visti come una gerarchia di cluster dove il nodo al più alto livello contiene tutti i dati, i quali vengono successivamente partizionati nei livelli più bassi dell'albero. I PCT combinano metodi di clustering e metodi di classificazione, in quanto dopo aver individuato insiemi che contengono oggetti simili tra loro (i cluster) viene associato un modello predittivo a ciascuno di essi che fornisce una predizione del valore di classe di nuove istanze le cui etichette di classe sono sconosciute.

In un lavoro di Vens et al [3] si propone uno dei più recenti sistemi di MLC gerarchica. L'algoritmo su cui si basa è molto simile agli algoritmi per la costruzione di alberi di decisione classici con la differenza che in una foglia non si ha una classe ma una gerarchia di classi rappresentata con un vettore binario che contiene uno 0 o un 1 a seconda che quella classe sia stata predetta o meno. L'euristica su cui si basa per la scelta del miglior test da applicare ad un nodo si basa su varianza delle etichette di classe rispetto all'insieme di training. La fase di predizione si basa su un concetto modificato di classe di maggioranza. Pertanto, nelle foglie troviamo un vettore contenente i valori medi per ciascuna classe che esprime, in modo probabilistico, l'appartenenza o meno di un esempio non visto a quella classe. A questo punto si fissa una soglia e si classifica sulla base dei valori di classe che superano la soglia. Questo lavoro non tiene conto del concetto di autocorrelazione.

Successivamente Cerri et al [4] propongono un algoritmo genetico per la risoluzione del problema HMC. In questo metodo l'antecedente di una regola di decisione evolve tramite una funzione di idoneità parziale nei confronti di regole con alta copertura di esempi, inoltre rimuove le istanze di training già coperte dalle regole già generate.

Specificatamente per il dominio biologico, Valentini [5] ha utilizzato un sistema di ensemble learning per la predizione dove le predizioni positive per un nodo influenzano transitivamente i suoi antenati nella gerarchia e le cui predizioni negative influenzano transitivamente i figli del nodo

corrente. Questo è uno dei primi approcci che considera la rete informativa, ma si limita a considerare la presenza o l'assenza delle interazioni, ignorando l'eventuale forza.

Una recente revisione dei lavori in letteratura ha mostrato che le tecniche che usano le reti PPI per la predizione della funzione delle proteine possono essere divise in due approcci maggiori. Il primo esplora direttamente gli schemi delle annotazioni dirette e inferisce la funzione di una proteina basata sulle sue connessioni nella rete. Il secondo si basa su moduli, identificando in una prima fase i moduli delle proteine correlate e successivamente annota ogni modulo.

Il primo approccio è utilizzato da Letovsky e Kasif [6] che utilizzano un modello basato sul campo aleatorio di Markov. Anche Kararoz et al [7] propone un metodo del primo approccio che vede la rete come un grafo, dove ogni nodo è un gene e un arco viene creato in base alla funzione dei nodi.

Il secondo approccio è utilizzato da Milenkovic e Przulj [8] che correlano la rete PPI alla struttura delle proteine complesse. Essi raggruppano le proteine considerando una topologia locale della rete e mostrano che questi gruppi di proteine che appartengono alla stessa proteina complessa, che svolgono la stessa funzione, sono localizzati nello stesso compartimento. Un altro lavoro che appartiene al secondo gruppo di approcci è quello di Borgwardt et al. [9], che utilizza i graph kernel che misurano la similarità tra grafi e apprendono un classificatore SVM per la predizione della funzione delle proteine, combinando informazioni sequenziali, strutturali e chimiche relative alle proteine.

In un lavoro più recente di Stojanova et al [10] si propone NHMC (Network Hierarchical Multi-Label Classification), un algoritmo basato su CLUS-HMC che, a differenza di quest'ultimo, tiene conto dell'autocorrelazione. In particolare, l'euristica utilizzata per la scelta del test da applicare ad ogni nodo consiste nella combinazione lineare della varianza e di una misura (basata sulla misura C di Geary) che indica l'autocorrelazione media. L'autocorrelazione è misurata su una rete informativa presa in input, considerata come background knowledge da sfruttare durante la fase di apprendimento. Questa rete è un grafo non orientato dove ogni nodo corrisponde ad un'istanza, ciascun arco indica la presenza di una interazione tra due istanze (autocorrelazione) ed il peso dell'arco rappresenta la forza di interazione. In questo lavoro, l'algoritmo proposto è applicato nel dominio della bioinformatica. In particolare, la rete informativa è rappresentata dalle PPI network (Protein to Protein Interaction network) e la predizione riguarda l'insieme di funzioni (organizzate gerarchicamente) che svolge un particolare gene.

Come detto in precedenza, anche nel dominio della text categorization la MLC gerarchica trova larga applicazione. Un lavoro che si occupa di text categorization gerarchica è quello di Malerba e Ceci del 2007 [11], che sfrutta l'organizzazione gerarchica delle etichette in tutte le fasi della classificazione. In particolare rappresenta i documenti a diversi livelli di astrazione in relazione al livello gerarchico per il quale si sta classificando il documento (le feature cambiano ad ogni livello della gerarchia). Si ricercano le categorie in modo top-down nella gerarchia e una volta trovata la categoria più generale, il classificatore (che può essere basato su naive Bayes, SVM o su centroidi) fornisce un punteggio per ogni sottocategoria, che esprime la forza di appartenenza di quel documento a quella sottocategoria. Successivamente si selezionano i punteggi che superano il valore soglia associato precedentemente a quella sottocategoria e tra quelli che superano la soglia si prende il migliore per classificare il documento in quella sottocategoria. La ricerca continua in modo ricorsivo sino a che non si ottengono punteggi tali che nessuno di questi superi la soglia per ogni sottocategoria. Se la ricerca si ferma alla radice il documento è considerato come non classificato. Questo approccio, tuttavia, non è in grado di associare ad ogni documento più etichette su concetti differenti e non è in grado di sfruttare la struttura a rete delle pagine ipertestuali.

6) Approccio al problema

Sfruttando tecniche di apprendimento automatico, si intendono sintetizzare algoritmi di SOP investigando, in particolare, i seguenti aspetti:

- 1) La rappresentazione e l'utilizzo di dati organizzati in reti informative in presenza di autocorrelazione, in grado di descrivere direttamente un nodo della rete oppure tramite le features del vicinato di un nodo, che infine vengono aggregate al nodo stesso.
- 2) L'applicazione di metodi di SOP, tra cui: MLC, HMC e Classificazione Multidimensionale in combinazione con apprendimento supervisionato e semi-supervisionato, dando enfasi alla Classificazione Multidimensionale che rappresenta la forma più complessa di output strutturato tra quelle appena descritte.
- 3) La valutazione sperimentale degli approcci proposti, in modo da individuare la combinazione più promettente tra dati di input, output del sistema e metodo SOP utilizzato.

I metodi che si intendono investigare sono quelli del clustering predittivo, dove il clustering può anche essere multi-tipo. Questo approccio permette l'uso naturale del setting semi-supervisionato, anche in un contesto in cui l'autocorrelazione è rappresentata utilizzando informazione sui vicini.

7) Ricadute applicative

Le ricadute applicative del presente progetto sono da ricercare nei settori della biologia computazionale e della web document classification.

Nel campo della biologia computazionale le applicazioni di tali sistemi potranno soddisfare i seguenti obiettivi: la scoperta di correlazioni tra genotipo e fenotipo, e la possibilità di scoprire collegamenti tra la composizione di microbiomi e lo stato di salute e nutrizione dell'organismo che li ospita.

Allo scopo di trovare una correlazione tra i dati del genotipo e quelli del fenotipo, si considereranno sia i geni individuali che i genomi. A livello di geni individuali si considererà il task di predizione SOP della funzione dei geni, sfruttando la MLC gerarchica e le annotazioni dei geni in Gene Ontology. Considerando anche la conoscenza biologica a priori sotto forma di reti di interazioni gene-proteina il problema si estende da SOP a network SOP. A livello di genomi completi, si considererà il task di predizione delle proprietà fenotipiche di microorganismi (ad esempio: bisogno di ossigeno, resistenza a osmosi, resistenza a variazioni di temperatura, patogenicità di piante e animali) a partire dal repertorio di geni presente nel genoma. Questi fenotipi sono stati a lungo trattati separatamente ma sono di fatto interdipendenti tra loro, infatti: si considererà il task di predizione simultanea come un task di SOP, che sfrutta la struttura e le dipendenze delle variabili di output da predire per produrre un modello più accurato. Questo metodo potrebbe essere molto utile se applicato a genomi dove alcune delle annotazioni fenotipiche non sono disponibili. Le annotazioni fenotipiche sistematiche sono scarse, pertanto è giustificato e motivato l'utilizzo di approcci di apprendimento semi-supervisionato. Differenti microorganismi sono collegati filogeneticamente tramite l'ereditarietà e il fatto di considerare queste relazioni introduce l'aspetto a rete di questo task.

La composizione di un microbioma, cioè un insieme di microbi che vivono su o dentro un essere umano (o animale) ospite, sono stati trovati su una grande varietà di individui e sono correlati a variabili come l'età o la nutrizione dell'ospite. È di grande interesse poter collegare la composizione

del microbioma a particolari situazioni di salute, come diabete o obesità, per comprendere i meccanismi delle patologie ed escogitare trattamenti tramite la manipolazione dei microbiomi. Si potranno applicare i metodi sviluppati per correlare la composizione dei microbiomi (in termini di insiemi di specie e loro metagenomi, come ad esempio l'insieme dei geni e le loro proprietà) con lo stato di salute e di nutrizione dell'individuo ospite, costituendo così un output strutturato.

Per quanto riguarda la classificazione applicata agli oggetti di una rete informativa multi-tipo, ci sono approcci che trovano applicazione sempre in ambito biologico. Un esempio di applicazione riguarda la rete informativa eterogenea dove gli oggetti, e quindi i nodi, sono di tre tipi: geni, patologie e farmaci. Gli archi pesati rappresentano la forza di interazione che sussiste tra coppie oggetti. Tramite tecniche di clustering predittivo, gli oggetti vengono raggruppati in modo gerarchico e, ad ogni livello, viene appreso un classificatore che applica tecniche SOP su uno o più attributi di classe identificati per uno o tutti i tipi di oggetti. Se ad esempio si considerasse come attributo di classe per il tipo "gene" la funzione del gene, il sistema predirà la funzione di tutti i geni la cui etichetta è ignota. Per classificare un gene, si considerano tutti i cluster in cui il gene compare e si sfruttano i geni, per i quali è già nota la funzione, che confluiscono nei cluster in cui compare anche il gene di cui si vuole predire la funzione.

Altri tipi di SOP trovano applicazione nel campo della web document classification. Un esempio è la classificazione gerarchica multidimensionale che è molto diffusa nei siti di e-commerce perché risulta essere ideale nella trattazione del prodotto da vendere mediante vari punti di vista o esigenze. In questo modo l'utente è più libero, infatti è in grado di interrogare i dati seguendo la logica che più si avvicina al proprio punto di vista, scegliendo quindi la dimensione su cui effettuare la ricerca. Sistemi di questo tipo consentono la possibilità di percorsi multipli sulla base dei diversi bisogni informativi e delle differenti tipologie di utenti.

8) Riferimenti bibliografici

[1] Pavlidis P, Weston J, Cai J, Noble WS: Learning gene functional classifications from multiple data types. *J Comput Biol* 2002, 9(2):401–411.

[2] Pandey G, Kumar V, Steinbach M: Computational approaches for protein function prediction. Tech. Rep. TR 06-028, Department of Computer Science and Engineering, University of Minnesota, Twin Cities 2006.

[3] Vens C, Struyf J, Schietgat L, Džeroski S, Blockeel H: Decision trees for hierarchical multi-label classification. *Mach Learn* 2008, 73(2):185–214.

[4] Cerri R, Barros RC, de Carvalho ACPLF: A genetic algorithm for hierarchical multi-label classification. In *Proc. 27th Annual ACM Symposium on Applied Computing*. New York: ACM; 2012:250–255.

[5] Valentini G: True path rule hierarchical ensembles for genome-wide gene function prediction. *IEEE ACM Trans Comput Biol Bioinformatics* 2010, 8(3):832–847.

[6] Letovsky S, Kasif S: Predicting protein function from protein/protein interaction data: a probabilistic approach. *Bioinformatics* 2003, 19(suppl 1): i197—i204.

[7] Karaoz U, Murali TM, Letovsky S, Zheng Y, Ding C, Cantor CR, Kasif S: Whole-genome annotation by using evidence integration in functional-linkage networks. *Proc Natl Acad Sci USA* 2004, 101:2888–2893.

[8] Milenkovic T, Przulj N: Uncovering biological network function via graphlet degree signatures. *Cancer Inform* 2008, 6:257–273.

[9] Borgwardt KM, Ong CS, Schönauer S, Vishwanathan SVN, Smola AJ, Kriegel HP: Protein function prediction via graph kernels. *Bioinformatics* 2005, 21(suppl 1): i47—i56.

[10] Stojanova, D., Ceci, M., Malerba, D., & Dzeroski, S. (2013). Using PPI network autocorrelation in hierarchical multi-label classification trees for gene function prediction. *BMC Bioinformatics*, 14, 285.

[11] Ceci, M., & Malerba, D. (2007). Classifying web documents in a hierarchy of categories: a comprehensive study. *Journal of Intelligent Information Systems*, 28(1), 37-78.

9) Fasi del progetto

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

- **Attività 1A:** studio approfondito di aspetti teorico-formali dell'apprendimento automatico e dello sviluppo di sistemi per la scoperta di conoscenza dai dati;
- **Attività 1B:** approfondimento delle tematiche relative alle tecniche di classificazione multi-label applicate a dati testuali;
- **Attività 1C:** ricerca e studio di metodi per la classificazione con output strutturato di oggetti di tipo diverso, applicati a dati organizzati in reti informative;
- **Attività 1D:** partecipazione a scuole internazionali inerenti all'attività e agli obiettivi previsti.

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

- **Attività 2A:** confronto con l'attività svolta da gruppi di ricerca con obiettivi affini;
- **Attività 2B:** sintesi, progettazione e implementazione di metodi per la classificazione con output strutturato di oggetti di tipo diverso, applicati a dati organizzati in reti informative, che soddisfino gli obiettivi previsti;
- **Attività 2C:** valutazione dei metodi realizzati, confronto con approcci esistenti e pubblicazione dei risultati conseguiti in riviste e conferenze internazionali.

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

- **Attività 3A:** stage presso università straniera e confronto con l'attività svolta presso altri gruppi di ricerca con obiettivi affini;

- **Attività 3B:** affinamento dei metodi e realizzazione di caratteristiche specifiche per il dominio applicativo scelto;
- **Attività 3C:** analisi dei risultati sperimentali ottenuti sul particolare dominio applicativo scelto;
- **Attività 3D:** stesura della tesi di dottorato.

10) Valutazione dei risultati.

In letteratura esistono varie misure di qualità applicabili ai risultati ottenuti da algoritmi di classificazione. Esse sono:

- **Precision**

$$\pi = \frac{TP}{TP + FP}$$

- **Recall**

$$\rho = \frac{TP}{TP + FN}$$

- **FMeasure**

$$F = \frac{2\pi\rho}{\pi + \rho}$$

- **Accuracy**

$$ACC = \frac{(TP + TN)}{(TP + TN + FP + FN)}$$

- **Area Under the ROC Curve**

La curva ROC rappresenta sull'asse delle ascisse di un piano il False Positive Rate (FPR) e sull'asse delle ordinate il True Positive Rate (TPR), dove:

$$FPR = \frac{FP}{FP+TN} \quad \text{e} \quad TPR = \frac{TP}{TP+FN}$$

L'area al di sotto della curva ROC permette di confrontare più classificatori tenendo conto della variazione della soglia che si può scegliere per discriminare se classificare nella classe positiva o negativa.

Queste misure devono estese prima del loro utilizzo qualora il classificatore sia multiclasse, infatti in questi casi vengono utilizzate nella variante micro averaged e macro averaged. Inoltre a seguito delle tecniche di SOP i concetti di True Positive (TP), True Negative (TN), False Positive (FP) e False Negative (FN) saranno opportunamente ridefiniti.

11) Eventuali referenti esterni al Dipartimento

Durante gli studi, nonché durante la partecipazione a scuole estive inerenti agli obiettivi prefissati, saranno selezionati alcuni referenti stranieri, operanti presso Università della Comunità Europea, al fine di supportare il lavoro di stesura della tesi di dottorato. Possibili referenti esterni sono:

- Ulf Brefeld - Technische Universität Darmstadt, Department of Computer Science, Robert-Piloty-Gebäude, Hochschulstr. 10, 64289 Darmstadt, Germany