UNIVERSITÀ DEGLI STUDI DI BARI ALDO MORO

Dipartimento DI INFORMATICA

Dottorato di ricerca in Informatica
XXVII ciclo

Progetto di ricerca

Dottorando: Dott. Gianvito Pio
Tutor: Prof. Michelangelo Ceci

Coordinatore
Prof. Donato Malerba

Firma del dottorando _____

Firma del tutor _____

1) Titolo della ricerca:

Collective relational clustering

2) Area nella quale si inquadra la ricerca:

Apprendimento automatico e Data Mining

3) Obiettivi della ricerca

Il problema di individuare gruppi di oggetti simili a partire da un insieme dato di oggetti (clustering) è stato ampiamente affrontato in letteratura. Recentemente, a causa di nuove sfide poste da particolari domini applicativi, come l'analisi di dati biologici o di documenti testuali, l'attenzione si è focalizzata sulla sintesi di algoritmi di co-clustering (o biclustering), con l'obiettivo di raggruppare simultaneamente oggetti e feature (cioè in base a due dimensioni).

Volendo considerare il clustering di oggetti secondo più di due dimensioni, è possibile trarre vantaggio dalla ricerca condotta nell'ambito del Multi Relational Data Mining e, più in particolare, del clustering relazionale. Tuttavia, sebbene gli approcci tradizionali al clustering relazionale prendano in considerazione le relazioni che intercorrono tra oggetti di tipo diverso, essi prevedono il raggruppamento di oggetti di un solo tipo. Solo in alcuni recenti lavori sono stati proposti alcuni approcci preliminari al *collective relational clustering*, nei quali il raggruppamento coinvolge oggetti di tipo diverso. Tuttavia, in tali approcci non è possibile lavorare su basi di dati relazionali di qualsivoglia struttura o indicare i tipi di oggetti da raggruppare (target) e quelli da utilizzare solo per analizzare la similarità tra oggetti target (task-relevant). Obiettivo della ricerca sarà pertanto la sintesi di metodi per il collective relational clustering che superino tali limitazioni e che affrontino aspetti, attualmente poco esplorati, quali: 1) l'individuazione di cluster eventualmente sovrapposti (overlapping); 2) la strutturazione dei cluster in una gerarchia, in grado di esprimere relazioni a diverse granularità e consentire una migliore interpretabilità dei risultati; 3) l'autocorrelazione, ossia la correlazione incrociata di un attributo con sé stesso; 4) la stima automatica del numero di cluster dai dati; 5) la complessità computazionale.

4) Motivazioni della ricerca

Uno degli ambiti in cui sono stati applicati metodi di co-clustering è quello biologico, in particolare dell'analisi di dati di espressione genica. Sempre in campo biologico, in lavori più recenti è stata proposta l'applicazione di tecniche di co-clustering anche a predizioni di interazioni tra microRNA (miRNA) e messenger RNA (mRNA). In entrambi i casi, è necessario raggruppare simultaneamente oggetti di tipo diverso (geni e condizioni nel primo caso, miRNA e mRNA nel secondo).

Tuttavia, avendo la possibilità di raggruppare oggetti secondo più di due dimensioni si potrebbero prendere in considerazione ulteriori aspetti finora trascurati, come, ad esempio, il tessuto specifico (nell'analisi di espressioni geniche) o il particolare processo biologico e la relativa fase temporale in cui un miRNA si attacca a un mRNA (nell'analisi di predizioni di interazioni mRNA:miRNA). Risulterebbe inoltre interessante estendere l'analisi a dati relazionali (relativi, ad esempio, ai miRNA e mRNA, nonché alle famiglie di miRNA, ai cluster genici, ai pathway metabolici in cui sono coinvolti, ecc.), tenendo conto dell'autocorrelazione relazionale. In questo caso, la presenza di autocorrelazione consiste nella correlazione tra un attributo e sé stesso, sulla base delle relazioni che sussistono con oggetti dello stesso tipo o di tipi diversi. L'autocorrelazione offre un'opportunità unica per cogliere dipendenze altrimenti trascurate.

Data la caratteristica delle entità coinvolte di poter intervenire in più reti di co-regolazione, risulta necessario consentire la scoperta di cluster sovrapposti, in quanto, limitando l'appartenenza di ciascun oggetto a un singolo cluster, si individuerebbero reti di co-regolazione incomplete. In aggiunta, organizzare gerarchicamente i cluster individuati consentirebbe di analizzare affinità funzionali tra oggetti a diversi livelli di granularità.

Tuttavia, allo stato attuale, non vi sono metodi che presentino tutte le caratteristiche richieste (raggruppamento di oggetti secondo più di due dimensioni, analisi di dati relazionali di struttura arbitraria, gestione dell'autocorrelazione, scelta dei tipi di oggetti target e task-relevant, overlapping, organizzazione gerarchia dei cluster) contemporaneamente. Peraltro, la maggior parte dei metodi che presentano parte di tali caratteristiche richiede in input il numero di cluster da individuare, informazione che risulta spesso del tutto ignota. Quest'ultima problematica suggerisce, quindi, la necessità di studiare e definire nuove strategie per la stima automatica (dai dati stessi) del numero ottimale di cluster da estrarre, nonché rafforza la necessità di organizzare i cluster ottenuti secondo

una gerarchia che, oltre a consentire una migliore interpretazione dei risultati, permetta di mitigare tale problematica.

5) Stato dell'arte

Relativamente al problema del clustering in generale, in letteratura sono disponibili molti lavori pionieristici, quali K-Means [9], uno dei più conosciuti algoritmi basati su centroide, e DBSCAN [6], tra gli algoritmi più rilevanti basati su densità.

Soffermandosi su lavori più recenti e sugli obiettivi presi in considerazione, è possibile individuare due linee di ricerca principali, ossia quella relative al co-clustering e quella relativa al clustering relazionale.

Relativamente al co-clustering, è bene notare come tutti i lavori si pongano come obiettivo il raggruppamento simultaneo di oggetti e feature, mantenendo tuttavia una netta distinzione tra i due. In tali lavori, infatti, l'obiettivo principale risulta comunque essere il raggruppamento di oggetti in base alla similarità calcolata in funzione delle feature, sebbene vengano comunque evidenziate quali siano le feature che maggiormente contribuiscono a ciascun raggruppamento. Oggetti e feature risultano pertanto non intercambiabili.

In letteratura, le tecniche di co-clustering sono state sostanzialmente applicate a matrici termini-documenti [5, 8, 12] o, in ambito biologico, a matrici geni-condizioni, relative a dati di espressione genica [2, 13, 10, 3, 4].

In particolare, in [5], gli autori propongono un metodo basato sul partizionamento di grafi bipartiti, eseguito calcolando il secondo autovalore principale di una matrice ottenuta a partire dalla matrice di adiacenza del grafo. In [8, 12] il co-clustering è ottenuto calcolando una fattorizzazione (a 3 fattori) della matrice, ponendo vincoli sulla positività dei valori dei fattori e sull'ortogonalità dei vettori che costituiscono il primo e il terzo fattore (solo in [12]). Punti deboli di tali approcci riguardano l'elevata complessità computazionale, l'impossibilità di ottenere co-cluster sovrapposti e/o organizzati in una gerarchia e la necessità di specificare in input il numero di cluster desiderato, per ciascuna dimensione.

In ambito biologico, il lavoro presentato in [3] costituisce uno dei primi tentativi di applicazione di tecniche di co-clustering a dati di espressione genica. In particolare, gli autori propongono un approccio basato su ricerca euristica top-down, tramite la quale è possibile ottenere un co-cluster alla volta, partendo da un co-cluster contenente tutti gli oggetti e procedendo iterativamente all'elimi-

nazione o al reinserimento dell'oggetto riga o dell'oggetto colonna che comporta la minore diminuzione del valore di un'euristica, fino al soddisfacimento di un criterio di terminazione. Tuttavia, il metodo prevede l'introduzione di perturbazioni casuali nei dati al fine di consentire l'inclusione, nei co-cluster scoperti successivamente, di oggetti riga e oggetti colonna già considerati in precedenza.

In [10], geni e condizioni sono rappresentati tramite una matrice binaria che è ricorsivamente divisa in due sottomatrici più piccole (eventualmente sovrapposte). In [1] gli autori definiscono un co-cluster come una *order-preserving submatrix (OPSM)*, ossia come un gruppo di righe i cui valori inducono un ordine lineare per un sottoinsieme di colonne. Una sottomatrice è definita *order-preserving* se esiste una permutazione delle sue colonne tale che il valore in ciascuna riga è strettamente crescente.

In [4] gli autori propongono un meta algoritmo (ROCC) bottom-up che effettua il co-clustering in maniera gerarchica, fondendo iterativamente i co-cluster che risultano più vicini, in accordo a una funzione di distanza. Tuttavia, l'algoritmo restituisce solo l'insieme dei co-cluster ottenuti a seguito di tutte le fusioni (un solo livello).

In [2], gli autori propongono l'estrazione di co-cluster sovrapposti e organizzati in una gerarchia, seguendo tuttavia un approccio non-deterministico e limitando la strutturazione gerarchica a una sola dimensione e la sovrapposizione (overlapping) all'altra dimensione.

Relativamente ai metodi proposti in letteratura che lavorano su dati relazionali, in [7] gli autori propongono un approccio per il collective clustering, in cui viene effettuata una fattorizzazione di ciascuna matrice che rappresenta una relazione tra oggetti di due tipi diversi. La funzione obiettivo da minimizzare è globale, nel senso che considera la situazione di fattorizzazioni che condividono un fattore. L'approccio proposto in questo lavoro prevede il raggruppamento simultaneo di oggetti di tipo diverso, ma il risultato è costituito da un insieme di cluster (la cui cardinalità è fornita in input) per ciascun tipo di oggetti corredati da un grado di associazione tra coppie di cluster relativi a oggetti di tipo diverso.

In [11] gli autori propongono un algoritmo per il clustering simultaneo di oggetti di tipi diversi, in relazione tra loro. Alla base del metodo vi è la definizione di una funzione di similarità, che prende in considerazione sia la similarità in termini di feature che in termini di relazioni. Dati gli oggetti di due tipi T_1 e T_2 che risultano in relazione, il metodo consiste nel clustering degli oggetti

di tipo T_1, usando come spazio delle feature i cluster ottenuti per gli oggetti di tipo T_2 e viceversa (iterativamente) fino al soddisfacimento di un criterio di convergenza. Il metodo prende il nome di *RECOM: Reinforcement Clustering of Multi-Type Interrelated Data Objects*, proprio per la caratteristica di utilizzare il risultato del clustering su una dimensione come spazio delle feature del clustering sull'altra dimensione.

Punti deboli di entrambi gli approcci proposti in [11, 7] sono l'impossibilità di individuare cluster sovrapposti, l'assenza di un'organizzazione gerarchica del risultato, l'impossibilità di indicare quali siano gli oggetti target e quelli task-relevant e la necessità di indicare, per ciascun tipo di istanza da raggruppare, il numero desiderato di cluster. Inoltre, la struttura dei dati è limitata a relazioni binarie e l'autocorrelazione non è tenuta esplicitamente in considerazione.

6) Approccio al problema

Sfruttando tecniche di apprendimento relazionale, si intendono sintetizzare algoritmi di collective clustering, investigando, in particolare, i seguenti aspetti:

Dati relazionali e autocorrelazione. L'analisi di dati relazionali richiede lo studio e l'eventuale definizione di funzioni di similarità relazionali. Dati due oggetti i_1 e i_2, la funzione di similarità deve tenere conto non solo dei valori delle feature di i_1 e i_2, ma anche (ricorsivamente) delle feature degli oggetti (anche di altri tipi) con le quali esse sono in relazione. Integrando la possibilità di calcolare la similarità riconsiderando gli oggetti di tipi già analizzati ($T_1 \rightarrow T_2 \rightarrow T_1$), ponendo eventualmente limiti alla profondità massima della ricorsione, è possibile tener conto dell'autocorrelazione.

Concetti target e task-relevant. La possibilità di scegliere concetti target e task-relevant focalizzerà la ricerca dei cluster sugli oggetti identificati come target. La similarità tra oggetti sarà valutata in funzione sia di concetti target che task-relevant.

Overlapping. Sebbene si possano realizzare algoritmi di clustering che individuino intrinsecamente cluster eventualmente sovrapposti, è possibile anche prevedere l'individuazione degli oggetti che, ragionevolmente, possono appartenere e più cluster, a partire dal risultato di un algoritmo di clustering non-overlapping. A tal scopo, è possibile trarre vantaggio da metodi di apprendimento supervisionato. Infatti, costruendo un classificatore su ciascuna coppia di cluster C_i e C_j, è possibile assumere che gli oggetti non classificati corretta-

mente (sul medesimo training set) siano localizzati spazialmente al confine tra i due cluster e che dunque appartengano ad entrambi i cluster.

Organizzazione gerarchica. Dato un insieme di cluster di partenza, è possibile individuare una gerarchia di cluster di più alto livello (più generali) fondendo iterativamente coppie di cluster (una o più coppie per ciascun passo dell'iterazione). Criteri di preferenza per la fusione possono essere in funzione della similarità tra cluster o di altri criteri qualitativi specifici del dominio applicativo. Possibili criteri di stop possono essere il raggiungimento di un unico cluster o l'impossibilità di effettuare ulteriori fusioni garantendo un valore di qualità del cluster ottenuto superiore a una soglia data.

Scelta del numero di cluster. Per delegare la scelta del numero ottimale di cluster al sistema stesso, è possibile tener conto delle caratteristiche della distribuzione dei dati dell'intero dataset e cercare di riprodurle in ciascun cluster. Poiché l'individuazione del valore ottimo risulta computazionalmente impraticabile, ci si affiderà a tecniche di ricerca greedy guidate da euristiche che esprimano l'aderenza dei cluster (in fase di definizione) alla distribuzione dei dati dell'intero dataset.

7) Ricadute applicative

L'uso di tecniche di clustering può risultare utile in numerosi ambiti applicativi, tra i quali:

- **marketing**, per l'individuazione di gruppi di clienti simili (profili), al fine di indirizzare prodotti, servizi, promozioni, ecc. in base alle loro caratteristiche; per il raggruppamento di prodotti e servizi simili, per fornire raccomandazioni su altri prodotti o servizi da acquistare.

- **web**, per il raggruppamento di documenti simili, al fine di fornire raccomandazioni su altri documenti da visitare.

- **medicina e biologia**, per l'individuazione di geni che presentano caratteristiche funzionali simili; per l'individuazione di farmaci che presentano effetti simili; per il riposizionamento di farmaci (uso di un farmaco conosciuto in altri ambiti dal suo attuale utilizzo), ecc.

Gli approcci tradizionali tuttavia, a causa delle limitazioni conosciute, in molti casi forniscono risultati poco significativi e/o poco interpretabili e/o ottenuti a partire da un sottoinsieme ridotto dei dati realmente a disposizione.

Il raggiungimento degli obiettivi preposti porterà dunque ad un sostanziale accrescimento delle potenzialità di utilizzo, anche in ambiti finora ritenuti di secondo piano.

Come ambito applicativo di riferimento, si prenderà in considerazione quello biologico, soffermandosi sull'analisi delle predizioni di interazioni tra microRNA (miRNA) e messenger RNA (mRNA). I miRNA sono piccole molecole di acido ribonucleico (RNA), che fungono da regolatori post-trascrizionali attaccandosi a sequenze complementari di mRNA. In particolare, sono stati individuati diversi ruoli nella regolazione negativa e possibili coinvolgimenti nella regolazione positiva. Inoltre, controllando la produzione delle proteine dei geni, i miRNA risultano coinvolti in molti processi biologici e controllano numerosi pathway metabolici. L'obiettivo sarà dunque quello di individuare gruppi di miRNA che manifestano comportamenti simili nei confronti dei medesimi mRNA, analizzando dati relazionali, relativi a famiglie, cluster genici, pathway metabolici, ecc. e considerando, ove significativo, il raggruppamento secondo ulteriori fattori, quali i processi biologici e le fasi temporali degli stessi in cui ciascun miRNA si attacca a particolari mRNA.

L'importanza dei risultati del clustering, in questo caso, è data dalla possibilità di concentrare le sperimentazioni di laboratorio sulle unità di analisi che evidenziano caratteristiche interessanti o precedentemente sconosciute nel risultato del clustering, riducendo drasticamente i costi necessari rispetto ad un'analisi esaustiva.

8) Riferimenti bibliografici

[1] Amir Ben-Dor, Benny Chor, Richard Karp, and Zohar Yakhini. Discovering local structure in gene expression data: the order-preserving submatrix problem. In *Proc. of RECOMB '02*, pages 49–57, 2002.

[2] José Caldas and Samuel Kaski. Hierarchical generative biclustering for microrna expression analysis. In *Research in Computational Molecular Biology*, volume 6044 of *LNCS*, pages 65–79. 2010.

[3] Yizong Cheng and George M. Church. Biclustering of Expression Data. In *Proc. of ISMB'00*, pages 93–103, 2000.

[4] Meghana Deodhar, Gunjan Gupta, Joydeep Ghosh, Hyuk Cho, and Inderjit S. Dhillon. A scalable framework for discovering coherent co-clusters in noisy data. In *Proc. of ICML'09*, page 31, 2009.

[5] I. S. Dhillon. Co-clustering documents and words using bipartite spectral graph partitioning. In *Proc. of SIGKDD'01*, pages 269–274, 2001.

[6] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *KDD*, pages 226–231, 1996.

[7] Bo Long, Zhongfei (Mark) Zhang, Xiaoyun Wú, and Philip S. Yu. Spectral clustering for multi-type relational data. In *Proceedings of the 23rd international conference on Machine learning*, ICML '06, pages 585–592, New York, NY, USA, 2006. ACM.

[8] Bo Long, Zhongfei (Mark) Zhang, and Philip S. Yu. Co-clustering by block value decomposition. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, KDD '05, pages 635–640, New York, NY, USA, 2005. ACM.

[9] J. B. MacQueen. Some methods for classification and analysis of multivariate observations. In L. M. Le Cam and J. Neyman, editors, *Proc. of the fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 281–297. University of California Press, 1967.

[10] A. Prelic, S. Bleuler, P. Zimmermann, A. Wille, P. Bühlmann, W. Gruissem, L. Hennig, L. Thiele, and E. Zitzler. A systematic comparison and evaluation of biclustering methods for gene expression data. *Bioinformatics*, 22(9):1122–1129, 2006.

[11] Jidong Wang, Huajun Zeng, Zheng Chen, Hongjun Lu, Li Tao, and Wei-Ying Ma. Recom: reinforcement clustering of multi-type interrelated data objects. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, SIGIR '03, pages 274–281, New York, NY, USA, 2003. ACM.

[12] Jiho Yoo and Seungjin Choi. Orthogonal nonnegative matrix tri-factorization for co-clustering: Multiplicative updates on Stiefel manifolds. *Inf. Process. Manage.*, 46:559–570, 2010.

[13] Sungroh Yoon, Luca Benini, and Giovanni De Micheli. Co-clustering: a versatile tool for data analysis in biomedical informatics. *IEEE Trans. on inf. technology in biomedicine*, 11(4):493–494, 2007.

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 clustering applicate a dati rappresentati tramite vettori di caratteristiche;

Attività 1C: ricerca e studio di metodi recenti per il clustering simultaneo di oggetti di tipo diverso, applicati a dati organizzati in basi di dati relazionali;

Attività 1D: partecipazione a scuole internazionali e conferenze su argomenti 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 il clustering simultaneo di oggetti di tipi diversi a partire da dati relazionali, 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.

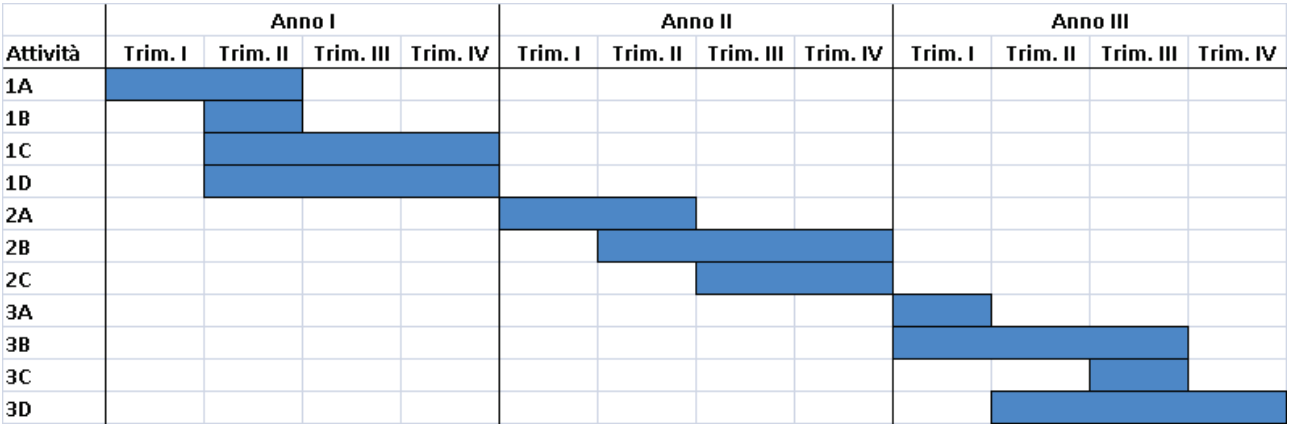
Anno 3°: applicazione al dominio applicativo scelto e sviluppo della tesi di dottorato:

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

Attività 3B: raffinamento 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 numerose misure di qualità applicabili ai risultati ottenuti da algoritmi di clustering, sia supervisionate che non supervisionate. Le prime, mutuate direttamente da task di apprendimento supervisionato (es. classificazione), sono applicabili solo se si dispone di un risultato di riferimento (ground truth), con il quale confrontare il clustering ottenuto. Esse sono:

- **avgPrecision(C)** $= \frac{1}{|C|} \sum_{i=1}^{|C|} \frac{TP_i}{TP_i+FP_i}$
- **avgRecall(C)** $= \frac{1}{|C|} \sum_{i=1}^{|C|} \frac{TP_i}{TP_i+FN_i}$
- **avgFMeasure(C)** $= \frac{1}{|C|} \sum_{i=1}^{|C|} 2 * \frac{precision(C_i)*recall(C_i)}{precision(C_i)+recall(C_i)}$

Prima di poter valutare tali misure è necessario effettuare un mapping tra i cluster ottenuti e quelli di riferimento, valutando, ad esempio, la similarità tra coppie di cluster o il numero di oggetti in comune.

Tra le misure non supervisionate vi sono la distanza media intra-cluster (da minimizzare), tra oggetti del medesimo cluster, e la distanza media inter-cluster (da massimizzare), tra oggetti di cluster diversi. A partire da esse, esistono in letteratura alcuni indici che valutano entrambi gli aspetti, come, ad esempio:

- **Davies-Bouldin Index**: $DB(C) = \frac{1}{|C|} \sum_{i=1}^{|C|} max_{i \neq j} \left(\frac{\sigma_i + \sigma_j}{d(c_i, c_j)} \right)$,

dove σ_i è la distanza media tra gli oggetti del cluster C_i e $d(c_i, c_j)$ è la distanza tra i centroidi dei cluster C_i e C_j (c_i e c_j, rispettivamente).

E' evidente tuttavia come sia necessario definire nuove misure e/o adattare quelle attualmente presenti in letteratura, al fine di prendere in considerazione gli aspetti tipici del clustering simultaneo di più concetti target e dei dati relazionali. Ad esempio, per la valutazione della coesione intra-cluster relativa un insieme di cluster costituiti da oggetti di k tipi, individuati a partire da un'associazione k-aria, è possibile calcolare la compattezza media:

- **compactness(C,A)** $= \frac{1}{|C|} \sum_{i=1}^{|C|} \frac{\sum_{x_1 \in C_i^{(1)}} \sum_{x_2 \in C_i^{(2)}} \cdots \sum_{x_k \in C_i^{(k)}} A_{m_1(x_1), m_2(x_2), \ldots, m_k(x_k)}}{|C_i^{(1)}| * |C_i^{(2)}| * \ldots * |C_i^{(k)}|}$

dove A è la matrice di adiacenza a k dimensioni, della quale ciascuna cella esprime la forza della relazione degli oggetti appartenenti a una k-upla; $C_i^{(j)}$ è l'insieme di oggetti di j-esimo tipo appartenenti al cluster C_i; $m_i(x)$, $1 \leq i \leq k$, sono funzioni che mappano l'oggetto x al relativo indice della matrice A dell'i-esima dimensione.

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:

- Philip S. Yu - IBM Watson Research Center, 19 skyline Drive, Hawthorne, NY 10532
- Ross D. King - Department of Computer Science, Aberystwyth University, Ceredigion, UK