UNIVERSITÀ DEGLI STUDI DI BARI ALDO MORO

Dipartimento DI INFORMATICA

Dottorato di ricerca in Informatica
XXVII ciclo

Progetto di ricerca

Dottorando: Dott. Gabriella Casalino
Tutor: Prof. Corrado Mencar

Coordinatore
Prof. Donato Malerba

Firma del dottorando _____

Firma del tutor _____

1) Titolo della ricerca: Metodi computazionali per l'estrazione di caratteristiche semanticamente rilevanti da dati espressi in forma matriciale

2) Area nella quale si inquadra la ricerca:

Intelligenza computazionale

3) Obiettivi della ricerca

Obiettivo della ricerca che si intende svolgere è di analizzare e sviluppare metodi per la fattorizzazione matriciale nell'ambito dell'Intelligent Data Analysis (IDA), con particolare interesse alle fattorizzazioni matriciali non negative. In particolare si analizzeranno i fattori di maggiore criticità della fattorizzazione matriciale nell'IDA, come la definizione degli iper-parametri necessari alla fattorizzazione e l'iniezione di conoscenza del dominio nel processo di fattorizzazione. Ciò al fine di sfruttare la fattorizzazione matriciale per estrarre fattori latenti nei dati che siano semanticamente significativi e dunque utili ai fini di un'analisi intelligente dei dati. Nel dettaglio si indagheranno le problematiche legate alla scelta del rango della fattorizzazione (che determina il numero di fattori latenti) ed eventualmente si proporranno tecniche semi-automatiche per la sua determinazione, anche sfruttando la conoscenza a priori sul dominio. Si analizzeranno inoltre le tecniche per la fattorizzazione basate sulla conoscenza, di solito basate sull'introduzione di vincoli come l'ortogonalità o la sparsità, con l'obiettivo di proporre metodi per l'iniezione di forme di conoscenza più complesse nel processo di fattorizzazione. I metodi sviluppati saranno oggetto di validazione sperimentale su dati sintetici e provenienti da ambiti reali, al fine di verificare l'efficacia della fattorizzazione matriciale nell'IDA in termini di accuratezza e significatività semantica dei risultati.

4) Motivazioni della ricerca

Lo sviluppo tecnologico comporta la produzione di una massa notevole di dati che, se da un lato costituiscono una fonte preziosa di informazioni, dall'altro determinano una serie di problemi legati all'estrazione e alla valutazione di questa massa di conoscenza. I vari aspetti di tali problemi sono affrontati in diverse aree dell'Informatica, tra cui la "Intelligent Data Analysis", che si basa sul processo di scoperta di conoscenza dai dati (Knowledge Discovery from Data) enfatizzando in particolare il ruolo che l'analista riveste nell'ambito di tale attività.
Per effettuare l'analisi intelligente dei dati, l'analista ha a disposizione numerosi strumenti utili sia per la verifica di ipotesi sia per l'analisi esplorativa dei dati. In quest'ultimo caso si può ricorrere a metodi diversi; tra questi, la fattorizzazione matriciale/tensoriale risulta particolarmente utile quando i dati possono essere organizzati in forma di matrici o tensori. Esempi classici di dati di questo tipo sono rappresentati dai dataset che costituiscono le matrici termini/documento, le espressioni genetiche, etc.
In generale, i metodi di 'rank reduction', tra cui le fattorizzazioni matriciali/tensoriali, permettono di ricavare una più utile rappresentazione dei dati in esame riducendone la dimensionalità, e hanno riscosso notevole interesse nell'ultimo decennio. Essi, tra l'altro, sono in grado di estrarre fattori latenti che accrescono ulteriormente il contenuto informativo che un dataset può offrire. I metodi di fattorizzazione sono stati ampiamente studiati da un punto di vista prettamente algoritmico e applicati in aree quali l'information retrieval e l'image processing. Tuttavia, si rileva una mancanza di studi organici che ponga l'enfasi sulla semantica dei fattori latenti; ciò ostacola l'impiego delle fattorizzazioni matriciali nell'ambito dell'Intelligent Data Analysis. Pertanto un approfondimento di questo tipo di tematiche risulta utile per produrre una migliore comprensione dei risultati ottenuti dall'applicazione di questi metodi e per consentire un loro più proficuo utilizzo.

5) Stato dell'arte

L'ammontare di dati disponibili è cresciuto drammaticamente negli ultimi 50 anni, si assiste ad un overloading di dati che arrivano dalle fonti più disparate: dati numerici provenienti ad esempio da satelliti, o sensori di ogni tipo, e dati testuali, strutturati e non, provenienti da siti web, email, forum, newsgroup, archivi digitali privati e pubblici, ecc. Se da un lato questa abbondanza di informazioni è un valore aggiunto perché permette di avere sempre a disposizione tutto ciò di cui si necessita, dall'altro è un limite, infatti quando la quantità di dati disordinati diventa troppa si traduce in "non esistenza" poiché è impossibile accedervi [55]. Poiché l'accumulo di dati è inevitabile, e cresce in maniera esponenziale, è necessario studiare meccanismi che ne permettano l'elaborazione automatica; per questo motivo negli ultimi anni si è visto uno sforzo estensivo della ricerca nell'ambito di tecniche che permettano la gestione ed elaborazione automatica di grandi quantità di dati e la ricerca di informazioni utili, ma nascoste in esse. Il processo globale di analisi di grossi database, finalizzata ad estrarre della conoscenza nascosta, è noto come "Knowledge Discovery in Databases" (KDD). Il Data mining (DM) è il cuore del processo di KDD e comprende gli algoritmi e le tecniche per esplorare ed apprendere dai dati, scoprendo conoscenza. Per enfatizzare la necessità di un'analisi intelligente dei risultati ottenuti, si usa il termine Intelligent Data Analysis [56].
In contesti che prevedono l'analisi di dati espressi in forma matriciale, metodi di blind source separation (BSS) e generalized component analysis (GCA) comprendono un'ampia classe di algoritmi di apprendimento non supervisionato che cercano di scoprire la struttura nascosta nei dati, estrarre feature significative e trovare rappresentazioni utili dei dati. Essi hanno trovato importanti applicazioni in diverse aree dall'ingegneria alle neuroscienze [59]. L'obiettivo delle GCA è di trovare rappresentazioni strutturate, ridotte o gerarchiche, delle componenti presenti nei dati osservati che possano essere interpretati. Poiché i dati possono sempre essere interpretati in molti modi differenti è necessario l'uso di conoscenza a priori per determinare quali feature o proprietà

meglio rappresentano le componenti latenti; quindi l'uso efficiente ed efficace di tali strumenti dipende fortemente da questa conoscenza a priori che permette di estrarre e identificare componenti nascoste significative.
Nel contesto brevemente delineato, i metodi in grado di approssimare la matrice dei dati, al fine di ridurne la dimensionalità (rank reduction) e di rimuovere il rumore, rivestono un'importanza fondamentale [56]. Esempi classici di tali metodi sono la fattorizzazione QR [1, 2], la decomposizione a valori singolari (Singular Value Decomposition, SVD) [3,8], l'analisi delle componenti indipendenti (Independent Component Analysis, ICA) [5] e l'analisi delle componenti principali (Principal Component Analysis, PCA) [4]. Questi metodi forniscono dei fattori, i cui elementi sono normalmente numeri reali, in taluni ambiti applicativi la presenza di valori negativi nei fattori fa perdere l'interpretazione semantica delle informazioni estratte dalle matrici. Paatero in [6] propose le Positive Matrix Factorization (PFM), applicate nell'ambito delle scienze ambientali, in cui per la prima volta si impose alle componenti la restrizione di non-negatività. Tuttavia il successo di tali tecniche si ebbe nel 1997 quando Lee e Seung proposero la "Nonnegative Matrix Factorization" (NMF) [7], un nuovo metodo di rank reduction con vincolo di non-negatività imposto sia sulla matrice di dati originali sia sui fattori, che permette di rappresentare i dati originari mediante combinazioni lineari, puramente additive, di basi non negative (part-based representation). Le NMF costituiscono un possibile approccio nel simulare la percezione umana dell'intero come somma degli elementi che lo costituiscono. Questo tipo di comportamento nell'uomo emerge spontaneamente dalla struttura neuronale in cui le soglie di attivazione dei neuroni non possono essere negative giustificando, quindi, il vincolo di non-negatività imposto ai dati [NATURE]. Le NMF hanno ricevuto attenzioni crescenti dalla comunità di analisi dei dati [20] grazie alla loro caratteristica di descrivere bene le proprietà del mondo reale e le interazioni tra di esse. La NMF applicata a una matrice di dati produce una approssimazione della stessa in termini di prodotto tra due matrici: la matrice delle basi (W) e la matrice di codifica (H) [57]. Una NMF di una matrice di dati è ottenuta mediante la risoluzione di un problema di ottimizzazione non lineare di una specifica funzione d'errore. La maggior parte degli algoritmi NMF utilizzano due funzioni errore: la distanza Euclidea e la divergenza di Kullback-Leibler; tuttavia vi sono esempi di algoritmi che utlizzano altri tipi di funzioni come la divergenza di Bergman, la divergenza di Csiszar, l'alfa divergence, la divergenza di Young, ecc.[METTERE CITAZIONE]
Lee e Seung proposero il "multiplicative update algorithm" [9], un algoritmo iterativo basato su regole di update delle matrici fattori per risolvere numericamente la NMF come un problema di ottimizzazione basato sul gradiente discendente. Questo algoritmo è stato utilizzato come baseline per lo sviluppo di numerose varianti, che aggiungendo dei vincoli alle matrici fattori, o alla matrice dei dati, introducono conoscenza nel processo di fattorizzazione [13]. Gli esempi più applicati in letteratura riguardano i vincoli di sparsità [10] e di ortogonalità [11] imposti ai fattori. Una delle proprietà di maggior interesse delle NMF è di generare una rappresentazione sparsa dei dati, in cui la totalità delle informazioni presenti nel dataset originale viene codificata utilizzando poche componenti positive. Nella formulazione classica delle NMF, tuttavia, la sparsità costituisce un effetto collaterale, piuttosto che un reale obiettivo, per cui non è possibile avere controllo sul grado di sparsità della rappresentazione. Per tale motivo Hoyer [10] propose una misura di "sparseness" utilizzata per imporre la ricerca di soluzioni che presentino il grado di sparsità desiderato. Alla base del vincolo di sparsità vi è l'idea che modelli sparsi siano rappresentazioni più semplici ed efficienti del problema. Inoltre, in molte applicazioni la conoscenza a priori sui dati suggerisce sparsità e quindi è naturale riversarla nella fattorizzazione.
Un altro aspetto semantico, affrontato mediante l'utilizzo di vincoli sulle regole di update, è l'ortogonalità dei vettori coinvolti nella fattorizzazione. Differentemente dagli altri metodi di rank reduction, i vettori base appresi dalla NMF non sono ortogonali tra loro, ciò indica che i fattori latenti possono non essere ben separati. In [58] è stato proposto un algoritmo NMF che impone una "quasi ortogonalità" ai vettori base, in corrispondenza con l'idea che i fattori latenti siano il più distanti possibile gli uni dagli altri e quindi esclusivi, pur preservando la loro non-negatività. In

particolare è stato dimostrato che quando il vincolo di ortogonalità è imposto alla matrice di codifica H, la NMF risulta essere equivalente all'algoritmo di clustering k-means [12].
Un caso particolare di dati sono le matrici binarie che rappresentano una relazione tra le feature e i dati rappresentati nella matrice iniziale. Zhang et al. in [14] propongono la Binary NMF, un algoritmo che fornisce uno schema naturale di normalizzazione delle matrici fattori a partire da una matrice binaria permettendo quindi di conservare la caratteristica relazionale.
Tutti gli algoritmi per il calcolo delle NMF sono iterativi e richiedono un'opportuna inizializzazione della matrice delle basi e delle codifiche, che influenza sia l'efficienza che l'efficacia: inizializzazioni povere provocano una convergenza lenta e una bassa riduzione dell'errore, producendo fattori latenti non abbastanza significativi.
Il problema dell'inizializzazione della NMF è stato ampiamente discusso in letteratura [15,60,62] e sono stati proposti differenti meccanismi di inizializzazione che permettano di ottenere un'alta riduzione dell'errore e una più veloce convergenza degli algoritmi NMF adottati. Tra questi si distinguono quelli basati su algoritmi di clustering sia hard, come lo spherical k-means[16] sia soft, come il Fuzzy C-Means (FCM) [17] e il subtractive clustering [19], quelli basati su schemi alternativi di approssimazione low rank [18], o basati su tecniche più avanzate come gli algoritmi genetici [19]. In generale strategie di inizializzazione complesse richiedono costi computazionali più alti, ma producono una più veloce riduzione dell'errore negli algoritmi NMF, e consentono un maggior determinismo nel processo di fattorizzazione.
Come già accennato in precedenza, il vincolo di non negatività e la rappresentazione part-based delle NMF ben si adattano a rappresentare dati del mondo reale; per questo motivo la letteratura relativa alle NMF spazia nei campi più svariati, tra cui il text mining, l'image processing, applicazioni in campo biologico che utilizzano la rappresentazione del DNA a microarray, in campo musicale, nell'e-learning, e contestualmente ad ogni problema specifico è stata assegnata una semantica ai fattori latenti estratti.
Sin dal primo articolo di Lee e Seung [7] le NMF sono state applicate nell'ambito del text-mining per estrarre concetti chiave (topic), da una collezione di documenti testuali memorizzati come colonne di una matrice termini-documenti [26]. In questo contesto i fattori latenti presenti nei dati sono un'insieme di parole chiave, estratte dal vocabolario dei documenti, che permettono di definire delle categorie semantiche in cui ciascun documento ricade. Quindi i documenti correlati ad un particolare "topic" possono essere raggruppati ed etichettati mediante il sottoinsieme di termini corrispondente al sottospazio vettoriale del vocabolario a cui appartengono [27]. Questa caratteristica ha permesso l'utilizzo delle NMF in diverse applicazioni reali come il riconoscimento di spam [23], l'information filtering basato sulle preferenze dell'utente [22,41], la classificazione di documenti [38], la creazione automatica di riassunti [25], il co-clustering [24], etc.
Un altro ambito in cui le NMF sono state utilizzate sin dal principio è il riconoscimento di volti. I dati sono rappresentati mediante una matrice pixel-immagini, le cui colonne sono linearizzazioni di immagini descritte mediante valori non negativi di pixel. Nel caso del face-recognition [36] i fattori latenti estratti dalla NMF sono caratteristiche fisiche semanticamente rilevanti e discriminanti, come ad esempio il naso, gli occhi, la bocca, gli zigomi nella stessa posizione in cui sono presenti nelle immagini originarie. A partire da questi risultati la NMF è stata applicata, con buoni risultati, nell'ambito dell'image processing a problemi di object detection [35,37] e di segmentazione [39,40], in cui i fattori latenti costituiscono parti significative delle immagini.
Negli ultimi anni si è assistito ad un crescente interesse per il "music information retrival". I dati che si vogliono analizzare sono la frequenza degli spettri musicali, al fine di eseguire automaticamente la separazione delle tracce di strumenti differenti o la trascrizione delle note. Molti sono gli articoli in letteratura che utilizzano le NFM sugli spettri sonori [42-50]. In questo contesto la matrice dei dati contiene il valore dello spettro in corrispondenza di una data frequenza e di un tempo fissato, i fattori latenti indicano per un singolo strumento la nota suonata e la matrice di codifica indica quando le note sono attive.
La NMF è stata applicata anche nel financial data mining, per analizzare i trend delle borse [51]. Si

è osservato che le fluttuazioni di prezzi in borsa non sono indipendenti l'una dall'altra ma dominate da numerosi fattori di fondo non osservabili, perciò si è pensato alla NMF come strumento automatico per identificare tali trend intrinseci nei dati. In questo ambito le colonne della matrice dei dati rappresentano i valori delle azioni all'evolversi del tempo. La NMF ha permesso di estrarre componenti intrinseche mediante le quali clusterizzare le azioni, e sorprendentemente si è osservato che azioni appartenenti allo stesso settore non sono state assegnate necessariamente allo stesso cluster aprendo spiragli nell'utilizzo della NMF come guida nella creazione di portfolii diversificati.
La NMF è stata applicata anche nel contesto di e-learning in cui alcuni dei dataset tipicamente disponibili sono rappresentabili attraverso score-matrix che registrano i punteggi ottenuti dagli studenti relativamente ai quesiti presenti nei test di verifica. A partire da una score matrix è possibile condurre un'analisi tesa a individuare i fattori latenti coinvolti nel processo di apprendimento. Le informazioni estratte rappresentano i building block di un modello cognitivo di apprendimento che trova corrispondenza in una particolare matrice, la cosiddetta question matrix (Q-matrix), che descrive le abilità necessarie affinché uno studente possa rispondere adeguatamente a questionari di valutazione. In [28] si utilizza la NMF come metodo automatico per la costruzione di una Q-matrix a partire da una score-matrix. Tuttavia questo non è l'unico esempio di applicazione della NMF nell'ambito della didattica; si rimanda a [29-33] per ulteriori applicazioni.
Una delle più recenti applicazioni della NMF è data dall'analisi di dati di tipo genomico. La rapida evoluzione della tecnologia microarray per l'espressione di geni ha fornito agli scienziati l'opportunità di osservare relazioni complesse tra i vari geni in un genoma, mediante la misura simultanea dei livelli di espressione di decine di migliaia di geni in esperimenti massivi. I microarray di espressioni di geni sono chip di silicio che misurano simultaneamente i livelli di espressione dell'mRNA di decine di migliaia di geni che tipicamente sono memorizzati in una matrice di dati, in cui ciascuna colonna rappresenta un gene e ogni riga un esempio o una condizione. Ogni valore della matrice rappresenta il livello di espressione di un gene nell'esempio corrispondente. La selezione di un sottoinsieme di geni discriminativi spesso aiuta a identificare geni che sono rilevanti come causa o conseguenza di una malattia, e possono essere usati come bio-marcatori per la diagnosi [52-54].
Questi sono alcuni esempi di ambiti in cui è stata maggiormente applicata la NMF, tuttavia la ricerca in tale campo è ancora molto aperta a diversi scenari.

6) Approccio al problema

Nel corso di questo progetto di ricerca si intende analizzare la semantica associata ai fattori latenti estratti dai dati adottando, come meccanismo di Intelligent Data Analysis, le fattorizzazioni matriciali (e in particolare, la fattorizzazioni non negative). Il problema di estrazione di caratteristiche semanticamente rilevanti da grandi matrici di dati è collocabile in un contesto globale di apprendimento non supervisionato, e quindi richiederà una profonda analisi delle diverse tecniche di clustering e dei loro meccanismi di funzionamento. La relazione tra alcune tecniche di clustering (quali il k-means) e le fattorizzazioni matriciali non negative, dovrà essere approfondita in modo da valutarne i punti di forza e le eventuali debolezze. Tale studio permetterà di identificare quali siano i meccanismi iterativi più idonei a inserire conoscenza durante il processo di fattorizzazione al fine di ottenere fattori latenti semanticamente rivelanti. In particolare, saranno analizzati i diversi algoritmi (proposti in letteratura) per calcolare le fattorizzazioni non negative al fine di verificare se l'introduzione di vincoli (nella formulazione delle regole di update) e l'utilizzo di appropriati meccanismi di inizializzazione possano iniettare conoscenza utile nel processo di fattorizzazione.
L'approfondimento della rappresentazione geometrica della fattorizzazioni di matrici e l'uso di tecniche di visualizzazione dei risultati di clustering potranno essere utilizzate per verificare la presenza di fattori latenti di raggruppamento nei dati. Anche lo studio dei metodi di valutazione

degli algoritmi di apprendimento potrà rivelarsi utile per validare la correttezza di eventuali ipotesi sui dati.
Acquisite le opportune conoscenze sulle tecniche di intelligent data analysis e sulle fattorizzazioni non negative, si provvederà alla fase di integrazione di queste in un'unica metodologia che consenta una più "intelligente" applicazione delle fattorizzazioni non negative. La metodologia ottenuta sarà sottoposta a verifiche sui diversi domini applicativi disponibili in letteratura in modo da valutare le proprietà semantiche e confrontare i risultati ottenuti con quelli noti dal punto di vista dell'efficienza e della significatività semantica.

7) Ricadute applicative

Come già evidenziato, le fattorizzazioni non negative in virtù delle intrinseche capacità di estrarre da dati una loro rappresentazione basata sulle parti (part-based), rappresentano un meccanismo idoneo a emulare le capacità percettive umane. In particolare, la rappresentazione part-based e la non negatività delle componenti che automaticamente possono essere ottenute applicando un metodo di fattorizzazione non negativa, permettono di estrarre automaticamente fattori latenti non noti a priori, ma implicitamente presenti nei dati. Le peculiarità dei metodi di fattorizzazione non negativa permettono il loro utilizzo per risolvere problemi di data mining relativi ai più disparati domini applicativi in cui i dati si possano rappresentare come matrici con elementi non-negativi. Esempio sono le immagini, i documenti testuali, le sequenze musicali, le sequenze di DNA, i dati di e-learning e quelli finanziari. Tuttavia, i risultati che si possono attualmente ottenere utilizzando tecniche di fattorizzazione necessitano dell'intervento di un esperto umano che in base a conoscenze a priori sia in grado di interpretare correttamente i risultati forniti da queste tecniche. La creazione di uno strumento automatico in grado, non solo di estrarre fattori latenti nascosti nei dati, ma di assegnare a tali fattori, in maniera quanto più automatica possibile, una corretta semantica sarebbe un passo avanti in tutti i campi di applicazione dell'intelligent data analysis (text clustering, recommender systems, E-learning data mining (EDM), financial data mining, image processing, signal processing, music information retrieval).

8) Riferimenti bibliografici

[1] Gene H. Golub and Charles F.Van Loan. Matrix computations (3rd ed.). Johns Hopkins Studies in Mathematical Sciences, 1996.
[2] Dario Bini, Milvio Capovani, Ornella Menchi. Motodi numerici per l' algebra lineare. Zanichelli, 1988
[3] Nash J. C., Compact Numerical Methods for Computers: Linear Algebra and Function Minimization, 2nd ed, Hilger, Bristol, 1990, 30-48.
[4] I. Joliffe. Principal Component Analysis. Springer-Verlag, 1986.
[5] Hyvarinen A., Fast and robust fixed-point algorithms for independent component analysis, IEEE Transactions on Neural Networks, 10, 3, 1999,626-634.
[6] P. Paatero and U. Tapper, "Positive matrix factorization: A nonnegative factor model with optimal utilization of error-estimates of data values," Environmetrics, vol. 5, no. 2, pp. 111–126, Jun 1994.
[7] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization." Nature, vol. 401, no. 6755, pp. 788–791, October 1999.
[8] G. Strang. Introduction to Linear Algebra (3rd ed.). Wellesley-Cambridge Press, 1998.
[9] Daniel D. Lee and Sebastian H. Seung. Algorithms for non-negative matrix factorization. In Proceedings of the Advances in Neural Information Processing Systems Conference, volume 13, pages 556–562. MIT Press, 2000.
[10] Patrik O. Hoyer. Non-negative matrix factorization with sparseness constraints. Journal of Machine Learning Research, 5:1457–1469, 2004.

[11] Wei Peng Haesun Park Chris Ding, Tao Li. Orthogonal nonnegative matrix trifactorizations for clustering. In 2th ACM SIGKDD international conference on Knowledge discovery and data mining, pages 126–135. ACM, 2006.
[12] Chris Ding, Xiaofeng He, and Horst D. Simon. On the equivalence of nonnegative matrix factorization and spectral clustering. In Proc. SIAM Data Mining Conf, pages 606–610, 2005.
[13] Murray Browne Michael W. Berry. Algorithms and applications for approximate nonnegative matrix factorization.
[14] Zhang, Z., Li, T., Ding, C., & Zhang, X. (2007). Binary Matrix Factorization with Applications. Seventh IEEE International Conference on Data Mining ICDM 2007, 391-400. Ieee.
[15] Del Buono, N., Lucarelli, M., Comparative studies on initializations for non negative matrix factorization algorithms, Tech. Rep. 17/10 (2010) Univ. Bari, Italy.
[16]Xue, Y., Tong, C. S., Chen, Y., Chen, W.-S., Clustering-based initialization for non-negative matrix factorization, Appl. Math. and Comp. 205 (2008) 525-536.
[17] Zhenga, Z., Yang, J., Initialization enhancer for non-negative matrix factorization, Eng. Appl. Art. Int. 20 (2007) 101-110.
[18] Boutsidis C.,G. E. , Svd based initialization: ahead start for nonnegative matrix factorization, Pattern Recognition 41 (4) (2008) 1350-1362.
[19] Casalino, G., Del Buono, N., and Mencar, C. (2011). Subtractive Initialization of Nonnegative Matrix Factorizations for Document Clustering. In Fuzzy Logic and Applications WILF 2011, A. M. Fanelli, W. Pedrycz, and A. Petrosino, eds. (Springer Berlin Heidelberg), pp. 188-195.
[20] Lee, D. D., Seung, S. H., Algorithms for non-negative matrix factorization, in: Proc. Adv. Neural Information Proc. Syst. Conf 13 (2000) 556-562.
[21] Andreas Janecek and Ying Tan. 2011. Using population based algorithms for initializing nonnegative matrix factorization. In Proceedings of the Second international conference on Advances in swarm intelligence - Volume Part II (ICSI'11), Ying Tan, Yuhui Shi, Yi Chai, and Guoyin Wang (Eds.), Vol. Part II. Springer-Verlag, Berlin, Heidelberg, 307-316.
[22] Takeru Yokoi , Hidekazu Yanagimoto , Sigeru Omatu. Information Filtering using Index Word Selection based on the Topics World Academy of Science, Engineering and Technology 50 2009
[23] Michael W. Berry Murray Browne. Email surveillance using nonnegative matrix factorization.
[24] Seungjin Choi Jiho Yoo. Orthogonal nonnegative matrix tri-factorization for co-clustering: Multiplicative updates on stiefel manifolds.
[25] Chan-Min Ahn Daeho Kim Ju-Hong Lee, Sun Park. Automatic generic document summarization based on non-negative matrix factorization.
[26]Michael W.Berry Robert J. Plemmons V.Paula Pauca, Farial Shahnaz. Text mining using non-negative matrix factorizations.
[27]Yihong Gong Wei Xu, Xin Liu. Document clustering based on non-negative matrix factorization.
[28] Desmarais M. C., Conditions for effectively deriving a q-matrix from data with non-negative matrix factorization, in Conati C., Ventura S., Calders T., Pechenizkiy M. (eds) Proceedings of the 4th International Conference on Educational Data Mining, 2011, 41-50.
[29] Lambropoulos N., Bakharia A., Gourdin A., Distributed leadership collaboration factors to support idea generation in computer-supported collaborative e-learning, Human Technology 7, 1, 2011, 72-102.
[30] Thai-Nghe N., Drumond L., Horvàth T., Krohn-Grimberghe A., Nanopoulos A., Schmidt-Thieme L., Factorization Techniques for Predicting Student Performance, in Santos O. C., Boticario J. G. (eds) Educational Recommender Systems and Technologies: Practices and Challenges, IGI Global, 2012, 129-153.
[31] Winters T., Payne T., What do students know?: an outcomes-based assessment system, in Anderson R., Fincher S. A., Guzdial M. (eds) Proceedings of the first international workshop on Computing education research, 2005,165-172.

[32] Yokoi, T., Yanagimoto, H., Omatu, S., Information filtering using index word selection based on the topics, World Academy of Science, Engineering and Technology 50, 2009
[33] Tong-Zhen Zhang, Rui-Min Shen, Hong-Tao Lu. "Using Non-Negative Matrix Factorization to Cluster Learners and Build Learning Communities". Chinese Journal of Electronics. Vol.22, No.2, pp207-211, 2011.
[34] David Guillamet and Jordi Vitri´a. Non-negative matrix factorization for face recognition. Lecture Notes in Computer Science, 2504:336–344, 2002.
[35] Weixiang Liu and Nanning Zheng. Non-negative matrix factorization based methods for object recognition. Pattern Recognition Letters, 25(14):893–897, October 2004.
[36] Bhavin J. Shastri and Martin D. Levine. Face recognition using localized features based on non-negative sparse coding. Machine Vision and Applications, 18:107–122, 2007.
[37] Daniel Soukup and Ivan Bajla. Robust object recognition under partial occlusions using nmf. Computational Intelligence and Neuroscience, vol. 2008:14 pages, 2008. Article ID 857453.
[38] Catarina Silva and Bernardete Ribeiro. 2009. Knowledge extraction with non-negative matrix factorization for text classification. In Proceedings of the 10th international conference on Intelligent data engineering and automated learning (IDEAL'09), Emilio Corchado and Hujun Yin (Eds.). Springer-Verlag, Berlin, Heidelberg, 300-308.
[39] Cosmin Lazar, Andrei Doncescu, and Nabil Kabbaj. 2010. Non Negative Matrix Factorisation clustering capabilities; application on multivariate image segmentation. Int. J. Bus. Intell. Data Min. 5, 3 (June 2010), 285-296.
[40] Jiayu Tang and Paul H. Lewis. 2008. Non-negative matrix factorisation for object class discovery and image auto-annotation. In Proceedings of the 2008 international conference on Content-based image and video retrieval (CIVR '08). ACM, New York, NY, USA, 105-112.
[41] Gang Chen, Fei Wang, and Changshui Zhang. 2009. Collaborative filtering using orthogonal nonnegative matrix tri-factorization. Inf. Process. Manage. 45, 3 (May 2009), 368-379.
[42] P. Smaragdis and J. Brown, Non-negative matrix factorization for polyphonic music transcription, in Applications of Signal Processing to Audio and Acoustics, IEEE Workshop on, Oct. 2003, pp. 177–180.
[43] H. Asari, "Non-negative matrix factorization: A possible way to learn sound dictionaries," Tony Zador Lab, Watson School of Biological Sciences, Cold Spring Harbor Laboratory, Tech. Rep., Aug 2005.
[44] M. Helén and T. Virtanen, "Separation of drums from polyphonic music using non-negative matrix factorization and support vector machine," in European Signal Processing Conference, Proceedings of (EUSIPCO), Sep 2005.
[45] M. N. Schmidt andM.Mørup, "Nonnegative matrix factor 2-d deconvolution for blind single channel source separation," in Independent Component Analysis and Blind Signal Separation, International Conference on (ICA), ser. Lecture Notes in Computer Science (LNCS). Springer, Apr 2006, vol. 3889, pp. 700–707.
[46]M. N. Schmidt and R. K. Olsson, Single-channel speech separation using sparse non-negative matrix factorization, in Spoken Language Proceesing, ISCA International Conference on (INTERSPEECH), 2006.
[47] S. A. Abdallah and M. D. Plumbley, Polyphonic transcription by non-negative sparse coding of power spectra, in Music Information Retrieval, International Conference on (ISMIR), Oct 2004, pp. 318–325.
[48] Y.-C. Cho, S. Choi, and S.-Y. Bang, Non-negative component parts of sound for classification, in Signal Processing and Information Technology, IEEE International Symposium on (ISSPIT), Dec 2003, pp. 633–636.
[49] Y.-C. Cho and S. Choi, Learning nonnegative features of spectro-temporal sounds for classification, Pattern Recognition Letters, vol. 26, no. 9, pp. 1327– 1336, Jul 2005.

[50] S. A. Raczy´nski, N. Ono, and S. Sagayama, Multipitch analysis with harmonic nonnegative matrix approximation, in Music Information Retrieval, International Conference on (ISMIR), Sep 2007.
[51] Konstantinos Drakakis, Scott Rickard, Ruairi de Frein, and Andrzej Cichocki. Analysis of financial data using non-negative matrix factorization. International Mathematical Forum, 3(38):1853–1870, 2008.
[52]H. Kim and H. Park. Improving molecular cancer class discovery through sparse non-negative matrix factorization. Bioinformatics, 21:3970–3975, 2005.
[53]Golub Mesirov Brunet, Tamayo. Metegenes and molecular pattern discovery using matrix factorization. In National Academy of Sciences, volume 101, pages 4164–4169, 2004.
[54]H. Kim and H. Park. Sparse non-negative matrix factorizations via alternating non-negativity-constrained least squares for microarray data analysis. Bioinformatics, 23:1495–1502, 2007.
[55] Computational Methods of Feature Selection Edited by Hiroshi Motoda and Huan Liu
Chapman and Hall/CRC 2007 Print ISBN: 978-1-58488-878-9
[56] Understanding Complex Datasets Data Mining with Matrix Decompositions David Skillicorn
Chapman and Hall/CRC 2007 Print ISBN: 978-1-58488-832-1
[56]Michael R. Berthold, Christian Borgelt, Frank Hppner, and Frank Klawonn. 2010. Guide to Intelligent Data Analysis: How to Intelligently Make Sense of Real Data (1st ed.). Springer Publishing Company, Incorporated.
[57] Cichocki, A., & Zdunek, R. (2008). Nonnegative matrix and tensor factorization. Signal Processing Magazine.
[58] Choi, S., Algorithms for orthogonal nonnegative matrix factorization, Proc. Intern. Joint Conf Neural Networks (2008)
[59] A. Cichocki and S. Amari. Adaptive Blind Signal and Image Processing. John Wiley & Sons Ltd, New York, 2003.
[60] R. Albright, J. Cox, D. Duling, A. N. Langville, and C. D. Meyer. Algorithms, initializa- tions, and convergence for the nonnegativematrix factorization. Technical report, NCSU Technical ReportMath 81706, 2006.
[61] Y.-D. Kim and S. Choi. A method of initialization for nonnegative matrix factorization. In Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP07), volume II, pages 537–540,Honolulu, Hawaii, USA, April 15–20 2007.
[62] A. N. Langville, C. D. Meyer, and R. Albright. Initializations for the nonnegativematrix factorization. In Proc. of the Twelfth ACM SIGKDD International Conference on Knowl- edge Discovery and DataMining, Philadelphia, USA, August 20–23 2006.
[63]O. Gillet and G. Richard, “Transcription and separation of drum signals from polyphonic music,” Audio, Speech, and Language Processing, IEEE Transactions on, vol. 16, no. 3, pp. 529–540, Mar 2008.

9) Fasi del progetto

Il progetto di ricerca sarà suddiviso in tre fasi:

1. Studio del problema;

2. Sviluppo della metodologia;

3. Valutazione dei risultati;

A ciascuna fase corrisponderanno le seguenti attività:

1. Studio della letteratura esistente inerente il problema e analisi degli strumenti disponibili;

2. Sviluppo di modelli e metodi legati alla metodologia proposta; implementazione di prototipi e strumenti che permettano di valutare la bontà dei modelli;

3. Raccolta di dataset benchmark disponibili in letteratura e compatibili i requisiti della metodologia proposta che permettano una efficace comparazione dei risultati. Valutazione dei modelli ottenuti attraverso adeguate metriche;

E' previsto, tra la prima e la seconda fase, un periodo di almeno tre mesi di studio presso una struttura di ricerca europea specializzata nella ricerca sulle fattorizzazioni matriciali.

Di seguito sono specificate le principali risorse da acquisire e la fase del progetto in cui si prevede saranno acquisite:

1. Algoritmi noti in letteratura;

2. Linguaggi di programmazione prototipale e general purpose;

3. Dataset benchmark (UCI Machine Learning Repository);

Ognuna delle tre fasi proposte verrà fatta coincidere con il rispettivo anno di dottorato. I risultati miliari che si presume vengano raggiunti sono i seguenti:

- Al termine del primo anno una survey inerente la letteratura disponibile sulla fattorizzazione matriciale non negativa applicata all'analisi dei dati;

- Durante il secondo anno lo sviluppo di una metodologia per l'applicazione della fattorizzazione matriciale non negativa nell'intelligent data analysis e sviluppo di relativi prototipi;

- Durante il terzo anno analisi sperimentale della metodologia e scrittura della tesi di dottorato;

10) Valutazione dei risultati.

La valutazione dei risultati è mirata a verificare la validità della metodologia proposta per la fattorizzazione matriciale non negativa nell'Intelligent Data Analysis. In particolare occorrerà verificare la significatività semantica dei risultati della fattorizzazione rispetto a tipici problemi di analisi dei dati. Questo richiede l'applicazione di metriche specifiche, nonché il feedback dell'analista nella valutazione dei risultati. In particolare, poiché la NMF può essere inteso come uno strumento di apprendimento non supervisionato verranno applicate le misure di validazione

interne ed esterne tipiche degli algoritmi di clustering, e saranno utilizzate rappresentazioni grafiche dei risultati che meglio evidenzino la presenza di eventuali raggruppamenti nei dati. Inoltre occorrerà valutare l'efficienza e l'efficacia delle tecniche di fattorizzazioni in termini di accuratezza dell'approssimazione e tempi di convergenza.

Al fine di valutare l'efficacia della fattorizzazione matriciale nell'analisi intelligente dei dati, si utilizzeranno dataset di diversa natura. In particolare si ricorrerà a dataset sintetici, anche costruiti ad hoc, al fine di verificare la capacità di tali metodi di estrarre fattori latenti attesi. Solo dopo questo tipo di validazione sperimentale si potrà ricorrere a dataset provenienti da problemi reali per una valutazione complessiva. Per rendere confrontabili e ripetibili le sperimentazioni che si effettueranno, verranno utilizzati dataset noti in letteratura e disponibili sul web. Nell'eventualità di dover testare che particolari caratteristiche semantiche affiorino dal processo di fattorizzazione, sarà possibile produrre dei dataset ad hoc che soddisfino le proprietà desiderate.

11) Eventuali referenti esterni al Dipartimento

Pierre Antoine Absil, Associate Professor Department of Mathematical Engineering ICTEAM Institute & Ecole Polytechnique de Louvain Université catholique de Louvain